# Viitekehys sosiaali- ja terveysalan ennakointimallien suorityskyvyn arviointiin

**Milja Asikainen**

**Perustieteiden korkeakoulu**

Diplomityö, joka on jätetty opinnäytteenä tarkastettavaksi diplomi-insinöörin tutkintoa varten Espoossa 26.2.2018.

**Työn valvoja:**

Prof. Harri Ehtamo

**Työn ohjaajat:**

TkT Mikko Nuutinen

Apul.prof Paulus Torkki

Aalto-yliopisto
Perustieteiden
korkeakoulu

AALTO-YLIOPISTO
PERUSTIETEIDEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

**Tekijä**: Milja Asikainen

**Työn nimi**: Viitekehys sosiaali- ja terveysalan ennakointimallien suorityskyvyn arviointiin

**Päivämäärä**: 26.2.2018 **Kieli**: Suomi **Sivumäärä**: 8+69

Matematiikan ja systeemianalyysin laitos

**Professuuri**: Systeemi- ja operaatiotutkimus

**Työn valvoja**: Prof. Harri Ehtamo

**Työn ohjaajat**: TkT Mikko Nuutinen, Apul.prof Paulus Torkki

Sosiaali- ja terveysalalla kerätään laajamittaisesti tietoa potilaiden terveydentilasta ja erilaisten resurssien käytöstä. Laajan aineiston avulla voidaan toteuttaa historiatiedon avulla tulevia tapahtumia ennakoivia malleja. Ennakoivan analytiikan käyttö sosiaali- ja terveysalalla mahdollistaa ennaltaehkäisevän ja yksilöllisemmän hoidon, jolloin voidaan sekä vähentää kuluja että parantaa ihmisten hyvinvointia. Ennakointimallien suorituskyky on tärkeää, mutta vaatimukset sille riippuvat sovelluskohteesta.

Työssä toteutetaan kirjallisuuskatsaus ja teemahaastatteluita ennakointimallien arvioinnista, joiden perusteella muodostetaan viitekehys laajentamalla ennakointimallien kehittämisen prosessia. Viitekehys ottaa suorituskyvyn arvioinnin huomioon sekä mallia suunniteltaessa, että rakennettua mallia validoitaessa. Suorituskykyä tarkastellaan ennustekyvyn, yleistyvyyden ja hyödyn näkökulmista.

Viitekehystä myös sovelletaan päivystystoimintaan liittyvään ennakointimalliin. Soveltamisen perusteella viitekehys toimii ennakointimallien kehitystyössä hyvänä muistilistana arvioinnissa tärkeistä asioista ja auttaa toteuttamaan hyödyllisiä malleja. Kun kokemusta ennakointimallien kehitystyöstä saadaan lisää, voi viitekehystä laajentaa koskemaan koko ennakointimallien kehitystyön prosessia tai tehdä ohjeistuksista yksityiskohtaisempia eri tyyppisiä malleja ajatellen.

**Avainsanat**: Ennakoiva analytiikka, ohjattu oppiminen, ennuste, arviointi, viitekehys, ennustekyky, yleistyvyys, hyöty, päivystysosasto, sosiaali- ja terveysala

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

**Author**: Milja Asikainen

**Title**: Framework for evaluating the performance of predictive models in healthcare and social services

**Date**: 26.2.2018 **Language**: Finnish **Number of pages**: 8+69

Department of Mathematics and Systems Analysis

**Professorship**: Systems and Operations Research

**Supervisor**: Prof. Harri Ehtamo

**Advisors**: D.Sc. (Tech.) Mikko Nuutinen, Assoc.prof. Paulus Torkki

In healthcare and social services data is collected on both patient wellbeing and patients' use of healthcare resources. Using this data, predictive models can be built. The use of predictive analytics in social services and healthcare allows for preventive and personalized actions which can lead to reduced costs and increased wellbeing. The performance of predictive models is essential. However, defining effective performance of these models depends on the context in which they are applied.

A literature review and thematic interviews about evaluation of predictive models is carried out. Based on the research a framework is formed by developing the process of building analytics models. The framework takes evaluation into account both in model planning and validation. It encompasses a variety of evaluation methods and divides the process of evaluation into parts. Model performance is viewed from the perspectives of accuracy, generalizability and utility.

The framework is applied to an emergency department related predictive model. The framework appears to be a useful checklist for model evaluation and it can help build useful predictive models. When more experience about development of predictive models is accumulated, the framework can be expanded to the whole process of model development and the guidelines can become more detailed for different kinds of models.

**Keywords**: Predictive analytics, supervised learning, prediction, validation, framework, accuracy, generalizability, utility, emergency department, social services and healthcare

# Esipuhe

Diplomityöni valmistui lopulta yllättävän nopeasti ja kivuttomasti, mistä kiitos kuuluu monille. Kiitos:

NHG, että pääsin keskittymään täysipainoisesti mielenkiintoiseen aiheeseen

Mikko ja Paulus tuesta, avusta, ohjauksesta ja ideoista koko työn ajan

Tomi, Ykä, Tommi ja Vesa asiantuntemuksen jakamisesta haastatteluissa

Emmi tarkasta oikoluvusta ja perusteellisista kommenteista

Kaikki muutkin NHG:n kollegat loistavasta seurasta töissä ja vapaa-ajalla

Harri sparrauksesta dipan kanssa ja tuesta läpi opintojeni

Diplomityö on teekkarin opintojen päätepiste, mutta ei välttämättä huipentuma. Vuosiin Aallossa on mahtunut luentoja ja laskareita, mutta kursseja opettavaisempaa on ollut opiskelijayhteisö ja sen mahdollisuudet. Kiitos FK ja Fyysikkospeksi, kun sain kantaa vastuuta ja ylittää itseni. Kiitos AYY ja rakkaat hallitukseni upeasta työyhteisöstä, jossa oli hyvä oppia. Kiitos Hankkijat ja Äpyn veijarit kaikista unohtumattomista ja jo unohtuneista hetkistä.

Kiitos rakkaat ystävät, jotka olette olleet tällä matkalla mukana. Ilman teitä olisin valmistunut vuosia aikaisemmin, mutta monta muistoa köyhempänä ja aika erilaisena ihmisenä.

*Tie on ollut pitkä ja (jalo)kivinen, eikä syyttä.*

Otaniemi, 14.2.2018

Milja M. Asikainen

# Sisällysluettelo

# Sanasto

| Lyhenne | Suomeksi | Englanniksi |
|---|---|---|
| | Binäärinen luokittelu | Binary classification |
| | Bootstrap-menetelmä | Bootstrap |
| | Ennakoiva analytiikka | Predictive analytics |
| | Ennustekyky | Accuracy |
| | Erottelu | Discrimination |
| | Hajontakuvio | Scatter plot |
| bias | Harhaisuus | Bias |
| | Hyöty | Utility |
| | Kalibraatiokuvio | Calibration plot |
| | Kalibraatiokuvion kulmakerroin | Calibration slope |
| | Kalibrointi | Calibration |
| MSE | Keskimääräinen neliövirhe | Mean squared error |
| | Kynnysarvo | Cut-off point, Threshold |
| | Luokittelija | Classifier |
| ACC | Luokittelun tarkkuus | Classification accuracy |
| | Moniluokkainen luokittelu | Multi-class classification |
| | Moninimikkeinen luokittelu | Multi-label classification |
| | Ohjattu oppiminen | Supervised learning |
| | Opetusjoukko | Training set |
| | Oppimiskäyrä | Learning curve |
| | Päätösanalyysi | Decision analysis |
| | Päätöspuu | Decision tree |
| | Ristiinvalidointi | Crossvalidation |
| AUC | ROC-käyrän alle jäävä pinta-ala | Area under the ROC curve |
| | Satunnainen osaotanta | Random subsampling |
| | Sekaannusmatriisi | Confusion matrix |
| | Siirrettävyys | Transportability |
| | Sisäinen validointi | Internal validation |
| | Spesifisyys | Specificity |
| | Suuren mittakaavan kalibrointi | Calibration-in-the-large |
| ROC-käyrä | Toimintaominaiskäyrä | ROC curve, reciever operating characteristic curve |

| Lyhenne | Suomeksi | Englanniksi |
|---|---|---|
| | Toistettavuus | Reproducibility |
| | Ulkoinen validointi | External validation |
| | Validointijoukko | Validation set |
| | Yleistyvyys | Generalisability |
| | Ylisovittuminen | Overfitting |

# Symbolit

| Lyhenne | Suomeksi | Englanniksi |
|---|---|---|
| $p_E$ | Esiintyvyys | Prevalence |
| $\hat{f}$ | Ennakointimalli | Prediction model |
| $Y_i$ | Havaittu vastemuuttuja näytteelle $i$ | Observed output |
| $K$ | Luokkien määrä | Number of classes |
| $h(X)$ | Luottamusmitta | Confidence measure |
| $\hat{\theta}$ | Mittarin $\theta$ estimaatti | Estimate of $\theta$ |
| $N$ | Otoskoko | Sample size |
| $N_{TN}$ | Oikeiden negatiivisten määrä | True negatives |
| $N_{TP}$ | Oikeiden positiivisten määrä | True positives |
| $N_{FN}$ | Väärien negatiivisten määrä | False negatives |
| $N_{FP}$ | Väärien positiivisten määrä | False positives |
| $X_i$ | Syötemuuttujat näytteelle $i$ | Input variables |
| $\hat{Y}_i$ | Ennustettu vastemuuttuja näytteelle $i$ | Predicted output |
| $c$ | Yhteensopivuusindeksi | Concordance index, c-index |

# 1 Johdanto

Sosiaali- ja terveyspalveluissa kerätään enenevässä määrin tietoa potilaiden terveydentilasta sekä palveluiden ja resurssien käytöstä. Data on esimerkiksi käynti- ja toimenpidetietoina, potilaskertomuksina, tutkimustuloksina, kuvantamisdatana ja laskutustietoina. Kattava aineisto toimii pohjana kehittää päätöksentekoa ja toiminnan suunnittelua hyödyntäen uudenlaisia menetelmiä. Tilastollisilla ja matemaattisilla malleilla datasta on mahdollista löytää uudenlaisia yhteyksiä. Laajojen aineistojen ja kehittyneiden mallien avulla voidaan esimerkiksi tehdä tarkempia diagnooseja, valita yksilökohtaisesti tehokkaimmat hoidot tai löytää potilaat, joilla on korkea riski sairastua. Tieto voi tukea erilaisia sosiaali- ja terveysalan toimintoja, kuten kliinistä päätöksentekoa, resurssien allokointia tai palveluiden suunnittelua. [1]

Tässä diplomityössä keskitytään ennakoivaan analytiikkaan ja erityisesti ennakointimallien suorituskyvyn määrittämiseen. Ennakoiva analytiikka mahdollistaa tulevan ennustamisen oppimalla laajasta historiadatasta. Sen vahvuudet terveydenhuollossa liittyvät proaktiivisuuteen ja yksilöllistymiseen. Toiminnan painopistettä voidaan siirtää sairauksien hoidosta ennaltaehkäisyyn ja populaatiotason ratkaisuista yksilöllisempään suuntaan, mikä sekä säästää kuluja että parantaa ihmisten hyvinvointia. Rajalliset resurssit voidaan kohdistaa niille, joille niistä on eniten hyötyä. Leskelä ym. [2] mukaan 10 % väestöstä kerryttää 80 % sosiaali- ja terveydenhuollon kustannuksista. Erityisesti tässä ryhmässä voidaan saavuttaa suuria hyötyjä, jos riskipotilaita pystytään tunnistamaan ja hoitamaan yksilöllisemmin.

Kyse on ihmisten terveydestä ja hyvinvoinnista, joten mallien suorituskyky on tärkeää. Vaatimukset mallin suorituskyvylle riippuvat sovellusalueesta ja mallin ominaisuuksista. Meidän tietojemme mukaan mallien arviointiin ei ole esitetty jokaiseen sovellukseen soveltuvaa tapaa tai keinoja määrittää riittävä suorityskyky tietyllä sovellusalueella.

Tässä diplomityössä toteutetaan kirjallisuuskatsaus, jossa perehdytään erilaisiin menetelmiin arvioida ennakointimallien suorituskykyä. Lisäksi haastatellaan asiantuntijoita, joilla on kokemusta ennakointimallien kehityksestä sekä osaamista sosiaali- ja terveysalalta. Tavoitteena on perehtyä laajasti ennakointimallien arvioinnin eri näkökulmiin ja menetelmiin sekä erityisesti sosiaali- ja terveysalan ennakointimallien kehitystyössä huomioon otettaviin asioihin. Kirjallisuuskatsauksen ja haastattelujen perusteella määritetään viitekehys sosiaali- ja terveysalan ennakointimallien suorituskyvyn arviointiin. Viitekehys jakaa ennakointimallien arvioinnin osakokonaisuuksiin ja kokoaa yhteen menetelmiä, joiden avulla arviointi voidaan toteuttaa tietyllä sovellusalueella.

Mallin suorituskyvyn varmistaminen vaatii arvioinnin huomioimista sekä ennakointimallia suunniteltaessa että sitä validoitaessa. Suunnitteluvaiheessa on tärkeää ymmärtää tutkittava ilmiö niin hyvin, että siitä voidaan löytää merkittävimmät ongelmat sekä suunnitella niitä ratkova ennakointimalli. Lisäksi tulee määrittää tavoitteet ennustekyvylle ja hyödylle. Mallin toteuttamisen jälkeen tulee puolestaan varmistua, että malli toimii tavoitteiden mukaisesti ja sen avulla voidaan muuttaa toimintaa parempaan suuntaan. Viitekehys pyritään toteuttamaan niin, että sen käyttö ei ole liian työlästä käytännön kehitystyössä, mutta mallin suorituskyvystä

saadaan varmistus.

Tässä diplomityössä toteutettu viitekehys tarkastelee ennakointimallien suorituskykyä kolmesta näkökulmasta, jotka ovat ennustekyky, yleistyvyys ja hyöty. Ennustekyky kuvaa sitä, kuinka yhteneviä mallin ennustamat ja havaitut arvot ovat. Yleistyvyys tarkoittaa mallin ennustekykyä ryhmälle, joka eroaa sen opettamiseen käytetyistä näytteistä. Se on tärkeää, sillä mallien tavoitteena on ennustaa lopputulos uusille näytteille [3]. Ennustekykyä ja yleistyvyyttä tarkastelemalla ei vielä voida määrittää mallin käytön seurauksia. Hyvän mallin käyttämisen tulisi tuottaa hyötyä käyttäjälle. Hyöty tarkoittaa, että mallin tuottamien ennusteiden avulla voidaan muuttaa toimintaa parempaan suuntaan ja säästää resursseja. [4, 5, 6]

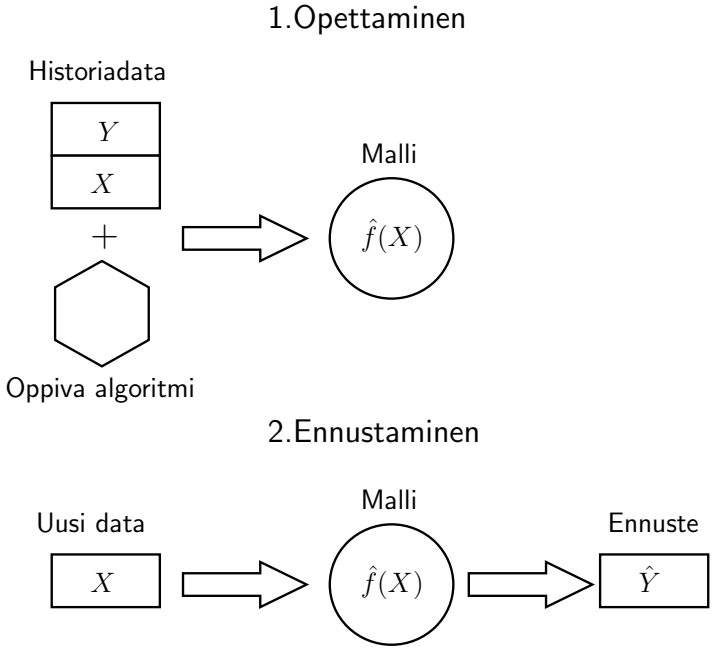
Diplomityö toteutetaan suomalaiselle yritykselle Nordic Healthcare Group (NHG), joka on erikoistunut sosiaali- ja terveyspalveluiden suunnitteluun ja kehittämiseen. Esimerkkejä yrityksen kehittämistä ennakoivan analytiikan sovelluksista ovat riskimittarit vanhusten kotihoitoon ja työterveydenhuoltoon liittyen. Vaikka yrityksessä tehdään ennakoivaa analytiikkaa sosiaali- ja terveysalalla, eivät mallit ole lääketieteellisiä vaan palveluprosessiin liittyviä. Yleensä vastemuuttujana on palvelutapahtuma, kuten päätyykö potilas osastolle, eikä niinkään kliininen tila. Kehitetyt mallit voidaan jakaa karkeasti kahteen luokkaan, riskityökaluihin ja kuormituksen ennakointityökaluihin. Riskityökaluissa keskiössä on yksilö. Yksilön palveluprosessia ennakoidaan, jotta löydetään ne yksilöt, joille kannattaa tehdä ennaltaehkäisevä toimenpide. Kuormituksen ennakointityökaluissa näkökulmana on yksikkö. Yksilötasolla ennustetut tulokset hoitoprosessista summataan, minkä pohjalta ennakoidaan potilaiden aiheuttamaa kuormitusta yksikkötasolla. Ennusteen pohjalta voidaan ohjata yksikkötason operatiivista toimintaa. Yrityksellä on käytössä prosessi ennakointimallien rakentamiseen, johon tässä diplomityössä määritetty arvioinnin viitekehys integroidaan. [7, 8]

Työ koostuu kahdeksasta luvusta. Luvussa 2 esitellään mitä ennakoiva analytiikka ja siinä käytettävät menetelmät ovat. Luvussa 3 esitellään erilaisia ennustekyvyn mittareita. Tämän jälkeen perehdytään mallin yleistyvyyteen ja siihen, miten voidaan laskea luotettavia arvioita mallin ennustekyvylle niin samassa kuin eri populaatiossa kuin millä malli on toteutettu (luku 4). Lisäksi käsitellään mallia rakentaessa käytetyn aineiston koon vaikutusta ennustekykyyn. Luvussa 5 pohditaan hyötyä ja esitellään erilaisia menetelmiä, joita voi soveltaa ennustemallin tuottaman hyödyn laskemiseen. Luvussa myös esitellään joitakin kirjallisuudesta löytyneitä ennustekykyä, yleistyvyyttä ja hyötyä huomioon ottavia keinoja mallien suorituskyvyn arviointiin.

Luvussa 6 esitellään prosessi ennakointimallien kehittämiseen ja laajennetaan sitä arvioinnin näkökulmasta toteutetulla viitekehyksellä. Viitekehystä sovelletaan päivystoimintaan liittyvään ennakointimalliin (luku 7). Lopulta viitekehyksen ansioita ja rajoituksia analysoidaan ja annetaan suosituksia lisätutkimuksille (luku 8).

# 2 Ennakointimallit

Ennakoivassa analytiikassa oppivaa algoritmia opetetaan historia-aineistoilla, joihin kuuluu syötemuuttujia $X$ ja vastemuuttuja Y. Näytteiden määrä, eli otoskoko, on $N$. Tavoitteena on tehdä ennakointimalli $\hat{f}$, joka pystyy saamaan aikaan mahdollisimman oikean vasteen $\hat{Y}$. Toisin sanoen, haluamme löytää mallin, joka parhaalla mahdollisella tavalla kuvaa datassa olevaa yhteyttä $Y = f(X)$. Mallin opettamisen jälkeen sitä käytetään uudelle datalle, jolle ennustetaan vastemuuttujan arvo, eli määritetään $\hat{Y}$. Prosessi mallin opettamiseen ja sillä ennustamiseen on havainnollistettu kuvassa 1.




Kuva 1: Ensin ennakointimalli muodostetaan opettamalla oppivaa algoritmia historiadatalla (ylempi kuva). Tämän jälkeen mallia käytetään syöttämällä siihen uutta dataa ja ennustamalla vastemuuttujan arvo (alempi kuva).[3]

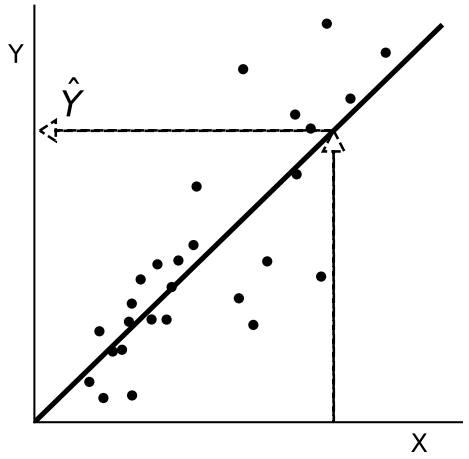
Steyerberg [9] esittelee prosessin ennakointimallin rakentamiseksi, arvioimiseksi ja käyttöönottamiseksi. Ensin data kerätään ja prosessoidaan käytettävään muotoon. Tähän vaiheeseen kuuluu muuttujien määrittely ja tarvittava muokkaaminen sekä puuttuvien arvojen käsittely. Toisessa vaiheessa valitaan oppiva algoritmi ja opetetaan malli. Oppivia algoritmeja on lukuisia ja usein opettaminen perustuu jonkin ennusteen ja opetusdatan välistä virhettä kuvaavan mittarin minimoimiseen mallin parametreja muuttamalla. Algoritmin valintaan vaikuttaa esimerkiksi vastemuuttujan tyyppi ja käytössä oleva data [10]. Luvussa 2.1 keskitytään vastemuuttujan tyyppiin ja luvussa 2.2 esitellään erilaisia oppivia algoritmeja. Algoritmeihin ja niiden ominaisuuksiin ja käyttökohteisiin ei keskitytä tarkemmin tässä diplomityössä.

Kolmannessa vaiheessa määritetään estimaatti mallin ennustekyvylle sekä tutkitaan mallin yleistyvyyttä ja sen käytöstä syntyvää hyötyä. Viimeisessä vaiheessa on olemassa arvioitu malli, joka tulee esittää sopivalla tavalla käyttäjille ja joka voidaan

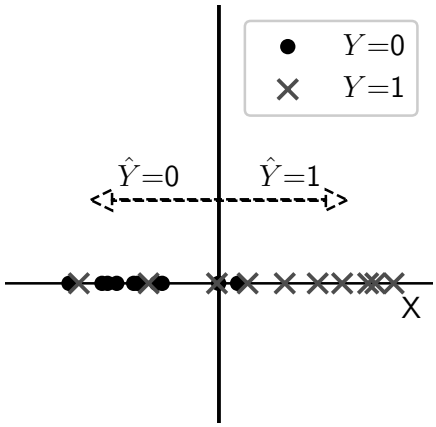
sen jälkeen ottaa käyttöön ja jota voidaan ruveta hyödyntämään uudelle datalle ennusteiden tekemiseen. Näin määriteltynä diplomityö keskittyy ennakointimallien rakentamisen prosessin vaiheeseen kolme, eli mallin arviointiin. Lisätietoa muista vaiheista löytyy kirjallisuudesta [9, 3].

## 2.1 Vastemuuttujan tyyppi

Ohjatussa oppimisessa vastemuuttuja voi olla joko jatkuva tai kategorinen. Kun vaste on jatkuva, kutsutaan ennakointiongelmaa regressioksi. Jos ennustetaan kategorista vastemuuttujaa, on kyse luokittelusta. [3, luku 1]




Kuva 2: Yksinkertaisessa lineaarisessa regressiossa havaituista syöte- ja vastemuuttujista $X, Y$ koostuvaan dataan sovitetaan suora, jonka avulla ennustetaan jatkuva vaste $\hat{Y}$ malliin syötetylle $X$:lle. Mukaillen [3]




Kuva 3: Yksinkertaisessa binäärisessä luokittelussa ennakointimalli on havaituista syöte- ja vastemuuttujista $X, Y$ laskettu syötemuuttujan kynnysarvo, jonka perusteella voidaan ennustaa näytteelle vaste $\hat{Y} \in 0, 1$. Mukaillen [3]

### Jatkuva vaste

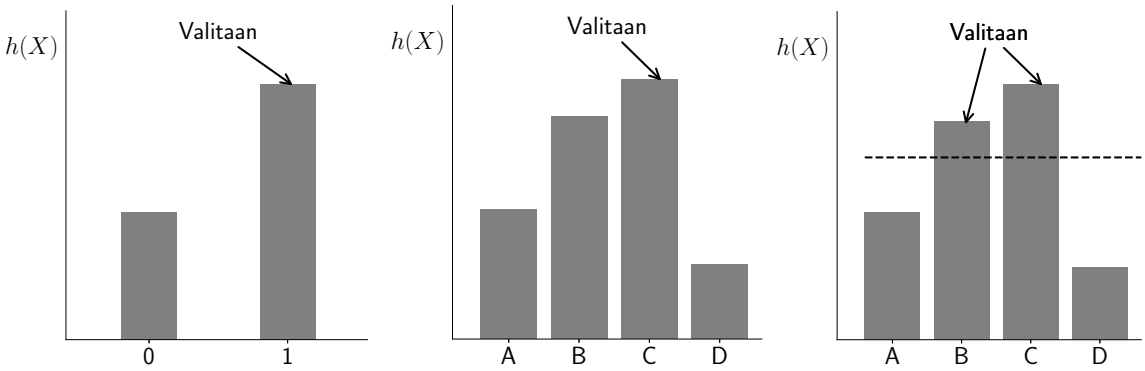
Regressiolla tarkoitetaan ohjatun oppimisen ongelmatyyppiä, jossa vastemuuttuja on jatkuva. Termiä ei kuitenkaan tule sekoittaa joukkoon regressiomenetelmiä, joilla voidaan menetelmästä riippuen ratkoa niin regressio- kuin luokitteluongelmia. Oppivasta algoritmista riippuen regression syötemuuttujat voivat olla joko jatkuvia tai kategorisia. Esimerkiksi syötemuuttujat voisivat olla ihmisen ikä, saadut diagnoosit ja edellisten lääkärikäyntien päivämäärät ja vastemuuttujana kuinka monta lääkärikäyntiä potilaalla on seuraavan vuoden aikana.

Yksinkertaisin menetelmä regressio-ongelmien ratkaisuun on lineaarinen regressio. Siinä vastemuuttujan arvo saadaan summaamalla yhteen parametreilla painotetut

syötemuuttujat $\hat{Y} = \beta_0 \sum_k \beta_k X_k$. Menetelmää on havainnollistettu yhden syötemuuttujan tapauksessa kuvassa 2. Parametrit $\beta$ estimoidaan datasta esimerkiksi minimoimalla summaa $RSS(\beta) = \sum(Y_i - \hat{Y}_i)^2$, jolloin menetelmää kutsutaan pienimmän neliösumman regressioksi (*least-squares regression*). Menetelmä ei ota kantaa mallin ennustekykyyn, vaan pyrkii löytämään parhaan lineaarisen sovitteen datalle. Lineaaristen regressiomallien hyviä puolia ovat yksinkertaisuus ja selkeästi tulkittava yhteys syöte- ja vastemuuttujien välillä. Wu ym. [11] pitävät todennäköisenä, että jopa 90% tosielämän ongelmista ratkaistaan kohtalaisen yksinkertaisilla lineaarisilla regressiomalleilla. Monet epälineaariset regressiomenetelmät ovat suoria yleistyksiä lineaarisista versioista. Menetelmiä regressio-ongelmien ratkaisuun esitellään luvussa 2.2. [3, luku 3]

### Kategorinen vaste

Luokittelussa mallin vastemuuttuja on kategorinen. Se voi olla binäärinen (*binary*), moniluokkainen *(multi-class*) tai moninimikkeinen (*multi-label*). Vastemuuttujatyyppien eroja havainnollistetaan kuvassa 4. Usein luokittelussa käytettävät oppivat algoritmit antavat pelkän luokan sijaan reaaliarvoisen todennäköisyyden tai muun luottamusmitan (*confidence measure*) näytteen kuulumisesta eri luokkiin. Merkitään tätä funktiolla $h(X)$ (moniluokkaisessa tapauksessa luokalle $j$: $h^j(X)$), jotta se erottuisi funktiosta $f(X)$, joka puolestaan antaa luokan. Usein luokitteluun käytetään sovelluksesta riippuvaa kynnysarvoa $p_T$ tai kynnysarvofunktiota (*thresholding function*) $t(X)$, jolloin $\hat{Y} = t(h(X))$. Kynnysarvo voi esimerkiksi olla 0.5, jolloin jos mallin ennuste näytteelle on korkeampi kuin 0.5, luokitellaan näyte luokkaan 1 ja muuten luokkaan 0. Malleja arvioitaessa kannattaa luokkien lisäksi hyödyntää myös luottamusmittoja $h(x)$, sillä ne tarjoavat mahdollisuuden monipuolisempaan, ennusteiden varmuuden huomioon ottavaan analyysiin.




(a) (b) (c)

Kuva 4: Binäärisessä luokittelussa (4a) vastemuuttujan luokka on toinen kahdesta, monen luokan tapauksessa (4b) vastemuuttuja on yksi $K$ luokasta ja moninimikkeisessä luokittelussa (4c) se voi olla useita $K$ luokasta, jolloin esimerkiksi valitaan kaikki tietyn kynnysarvon ylittävät nimikkeet. [3, 12]

Binäärisessä tapauksessa vastemuuttuja usein koodataan numeroilla nolla tai yksi, jossa yksi tarkoittaa, että tutkittava asia tapahtuu (onnistuu/sairastuu) ja nolla, että se ei tapahdu (epäonnistuu/ei sairastu). Monesti vasteista puhutaan positiivisena (1) tai negatiivisena (0). Mallia rakentaessa tavoitteena on tehdä yhteys $\hat{Y} = f(X)$, jossa $\hat{Y}_i \in [0, 1]$. Esimerkki yksinkertaisesta binäärisestä luokittelusta on esitetty kuvassa 3. [3, luku 2]

Moniluokkainen vastemuuttuja voi saada useita arvoja, jotka usein koodataan numeroin $Y_i \in 1, ..., K$, jossa K on luokkien määrä. Merkintänä $Y_i^j \in [0, 1]$ tarkoittaa näytteen $i$ kuulumista luokkaan $j$. Osa luokittelumenetelmistä yleistyy luonnollisesti useammalle luokalle. Niitä algoritmeja, jotka eivät suoraan yleisty, voidaan hyödyntää jakamalla luokittelu useisiin binäärisiin tehtäviin. Tämä voidaan tehdä esimerkiksi K:lla binäärisellä luokittelulla, jossa luokan $k$ havaintoja verrataan muiden $K - 1$ luokan havaintoihin tai tekemällä luokittelu jokaiselle parille erikseen, jolloin luokittelijoita tarvitaan yhteensä $\frac{K(K-1)}{2}$ kappaletta. Tehtävä voidaan myös jakaa hierarkkisesti. Tällöin luokat järjestetään puuksi, jonka solmuissa hyödynnetään binäärisiä luokittelijoita ja lehdet edustavat aina yhtä luokkaa. [13]

Kolmas vaihtoehto luokittelutehtävälle on se, että luokkia on useita, ja vastemuuttuja voi saada niitä samanaikaisesti arvokseen. Tällöin $Y_i \subseteq \{y_1, ...y_K\}$ on joukko nimikkeitä ja K on mahdollisten nimikkeiden kokonaismäärä. Moninimikeluokittelussa haasteena on yhdistelmien valtava määrä. Jos erilaisia nimikkeitä on esimerkiksi 20, on mahdollisia yhdistelmiä yli miljoona, jolloin niiden kaikkien tarkastelu erillisinä on vaikeaa. Osa menetelmistä tarkastelee jokaista nimikettä erikseen, mikä tekee menetelmistä yksinkertaisia ja tehokkaita. Toisaalta, niiden suorituskyky kärsii, kun nimikkeiden väliset riippuvuudet jätetään huomiotta. Jotkin menetelmät hyödyntävät pareittaisia riippuvuuksia, kuten järjestämistä relevanttiin ja epärelevanttiin nimikkeeseen. Joissain sovelluksissa nimikkeiden väliset riippuvuudet ovat kuitenkin tätä monimutkaisempia. Korkeamman kertaluokan riippuvuudet voidaan ottaa huomioon esimerkiksi määrittämällä yhteyksiä satunnaisten nimikkeiden osajoukkojen kesken. Nämä menetelmät huomioivat muita paremmin monimutkaiset yhteydet datassa, mutta ovat toisaalta laskennallisesti vaativampia ja huonommin skaalautuvia. [12]

Vastemuuttujan tyypin mukaan voidaan määritellä olevan myös hierarkkista luokittelua, jossa näyte voi kuulua vain yhteen luokkaan, mutta luokilla saattaa olla alaluokkia tai ne voivat kuulua suurempiin superluokkiin [14]. Tyyppinä on myös yhden luokan (*one-class*) luokittelu, jossa dataa on vain yhdestä luokasta, ja mallilla ennustetaan kuuluuko uusi havainto siihen [15]. Nämä vastemuuttujatyypit eivät kuulu tämän työn rajaukseen.

## 2.2 Oppivia algoritmeja

Erilaisia oppivia algoritmeja on lukuisia ja ne perustuvat hyvin erilaisiin menetelmiin. Osa menetelmistä, kuten lineaarinen regressio, perustuu tilastotieteeseen ja pyrkii vastemuuttujan ennustamisen ohella selittämään datasta löytyviä yhteyksiä. Esimerkiksi neuroverkot taas ovat enemmänkin 'black box' -malleja, jotka ovat niin monimutkaisia, että yhteyksiä on vaikeaa tai jopa mahdotonta ymmärtää. Osa menetelmistä soveltuu vain regressio-ongelmille, osa vain luokitteluun ja osa molempiin. Osa me-

netelmistä on täysin lineaarisia, jolloin ne perustuvat hypertasojen sovittamiseen, osa yhdistää useita paikallisesti lineaarisia malleja ja osa taas on lähtökohtaisesti epälineaarisia. Yleisesti, mitä monimutkaisempi ja epälineaarisempi malli, sitä paremmin se voi sovittua dataan. Toisaalta monimutkainen malli saattaa helpommin ylisovittua (*overfit*), jolloin se toistaa hyvinkin tarkasti siihen syötettyä dataa, mutta ei toimi uudella näytejoukolla. Kaikilla menetelmillä on omat rajoituksensa ja etunsa, eikä yhdenkään menetelmän voi sanoa olevan muita parempi kaikissa tilanteissa ja kaikille otosjoukoille. Yksi keino hyödyntää erilaisten menetelmien vahvuuksia on yhdistää useita perusmenetelmiä ylisovittumista vähentäväksi yhdistelmämalliksi (*ensemble learning*) [3, luku 16]. Näihin menetelmiin ei perehdytä tässä työssä. [11, 16]

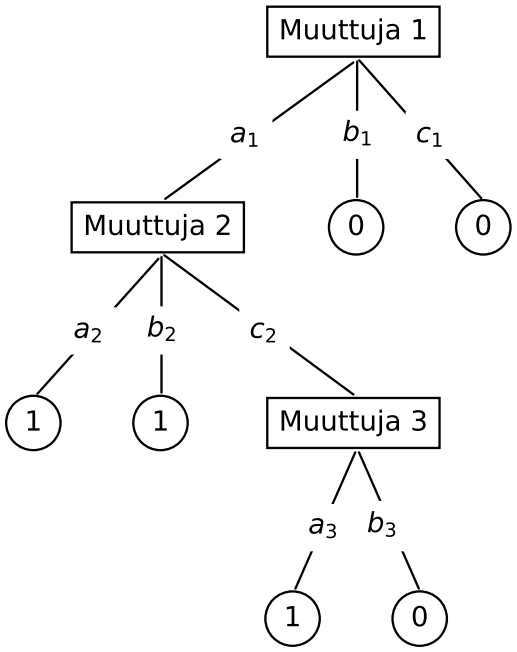
Kotsiantis ym. [16] jakaa menetelmiä viiteen tyyppiin: logiikkaan perustuviin tekniikoihin (*logic-based techniques*), perseptroneihin perustuviin tekniikoihin (*perceptron-based techniques*), tilastollisiin oppimisalgoritmeihin (statistical learning algorithms), näyteperustaisiin oppijoihin (*instance based learners*) ja tukivektorikoneisiin (*support vector machines*, SVM). Kuten hän kirjoittaa, tämä tyypittely ei ole kaikenkattava, mutta antaa kuitenkin yleiskuvaa tarjolla olevista menetelmistä. Alla on näiden tyyppien lisäksi esitelty erillisinä regressiomenetelmät niiden yleisyyden takia sekä annettu esimerkkejä erilaisista menetelmistä. Yhteenvetoa yleisistä menetelmistä ja niiden soveltumisesta regressio- ja luokitteluongelmiin löytyy taulukosta 3.

Taulukko 3: Esimerkkejä oppivista algoritmeista

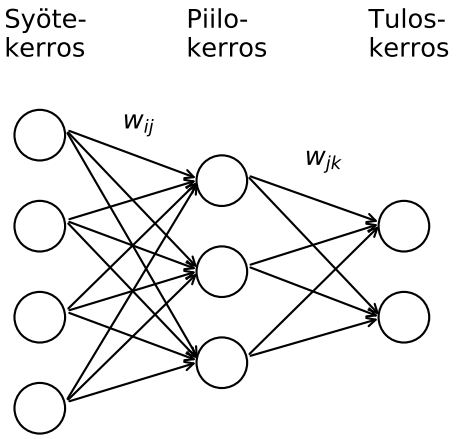
| Oppiva algoritmi | Regressio | Luokittelu | Algoritmityyppi |
|---|---|---|---|
| (Usean syötemuuttujan) lineaarinen regressio | x | | Regressio |
| Yleiset additiiviset mallit | x | | Regressio |
| Tukivektorikoneet | x | x | Tukivektorikoneet |
| Päätöspuut | x | x | Logiikkaan perustuvat tekniikat |
| k-lähimmän naapurin algoritmit | x | x | Näyteperusteiset menetelmät |
| Bayes-verkot | x | x | Tilastolliset oppimisalgoritmit |
| Neuroverkot | x | x | Perseptroneihin perustuvat tekniikat |
| (Usean syötemuuttujan) logistinen regressio | | x | Regressio |

*Regressiomenetelmistä* lineaarinen regressio esiteltiin luvussa 2.1. Siitä on lukuisia muunnelmia, joiden tavoitteena on mallin ymmärrettävyys, parempi ennustekyky tai ylisovittumisriskin pienentäminen. Muuttujat $X_j$ voivat olla standardisoituja, tai kuvata datan muuttujien muunnoksia ja niiden välisiä vuorovaikutuksia, esimerkiksi $X_3 = X_1X_2$ tai $X_2 = X_1^2$, kuitenkin kertoimien ollen lineaarisesti riippuvia paramet-

reista $\beta$. Opetusvaiheessa liitetään usein lisätermi, joka minimoi parametrien kokoa samalla kun malli sovitetaan dataan. Esimerkkejä tällaisista kutistamismenetelmistä (*shrinkage*) ovat ridge ja lasso. Muuttujista voidaan myös etsiä tehokas osajoukko esimerkiksi askeltava valinta -menetelmällä (*stepwise selection*). Reaalimaailmassa datassa olevat yhteydet ovat usein epälineaarisia, jolloin lineaariset mallit eivät pysty takaamaan parasta tulosta. Erilaiset additiiviset mallit tuovat joustavuutta lineaarisiin regressiomenetelmiin verrattuna säilyttäen tulkittavuutta. Yleistetyt additiiviset mallit (*Generalized Additive Models*) rakentuvat vastaavasti kuin lineaarinen regressiomalli, mutta vakiokertoimet $\beta$ korvataan sileillä funktioilla, jotka estimoidaan mallia sovitettaessa. Muita regressiomenetelmiä ovat esimerkiksi logistinen regressio, Poisson regressio ja polynominen regressio. [3]




Kuva 5: Päätöspuussa näytteet jaetaan solmuihin syötemuuttujien arvoihin perustuvien päätössääntöjen perusteella. Mukaillen [16]



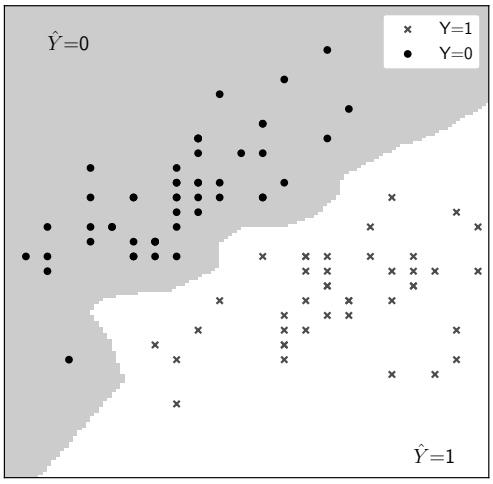

Kuva 6: Neuroverkoissa on toisiinsa painotetusti kytkettyjä laskentayksiköitä, joissa syötteet summataan ja lähetetään eteenpäin. Mukaillen [16]

*Logiikkaan perustuvat tekniikat* ovat symbolisia menetelmiä, joiden algoritmit pystytään esittämään sanoitettuina kaavioina. Esimerkkejä niistä ovat päätöspuut (*decision trees*) ja sääntöpohjaiset (*rule-based*) menetelmät. Ne perustuvat rekursiiviseen osiontiin, jossa muodostetaan lukuisia yksinkertaisia päätössääntöjä. Päätössääntöjen avulla näyte voidaan määrittää johonkin luokkaan tai voidaan sille laskea arvo. Esimerkki päätöspuusta on kuvassa 5.

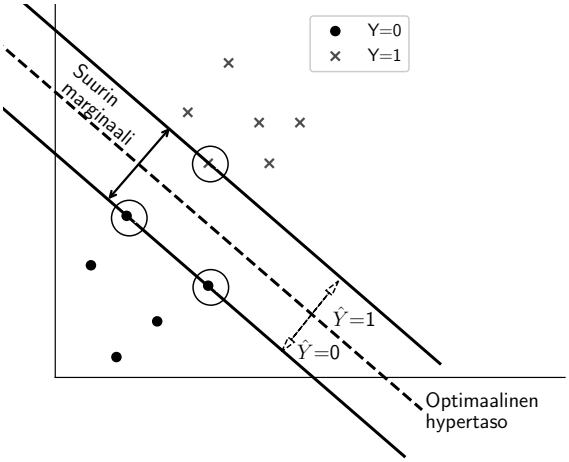
*Perseptroneihin perustuvat tekniikat*, kuten neuroverkot (*neural networks*), hyödyntävät toisiinsa kytkettyjä laskentayksiköitä. Kytkennöillä on painot, joiden avulla yksikköön saapuvat signaalit summataan, ja summasta lasketaan aktivaatiofunktion perusteella yksikön ulostulo. Tiedonkäsittely-yksiköitä voi olla lukuisia rinnakkain ja

kerroksittain, joten rakenteista voi tulla hyvinkin monimutkaisia. Opetettaessa mallia kytkentöjen painoja muunnellaan. Esimerkki neuroverkosta löytyy kuvasta 6. Binäärisessä luokittelussa usein hyödynnettävä logistinen regressio on tällä määritelmällä yksi laskentayksikkö. [16]

*Tilastollisissa oppimisalgoritmeissa* on eksplisiittinen tilastollinen malli, joka kertoo todennäköisyyden että näyte kuuluu johonkin luokkaan. Naiivi Bayes luokittelija on esimerikki tilastollisesta oppimisalgoritmista. Siinä tehdään oletus, jonka mukaan syötemuuttujat ovat riippumattomia, eli $P(X = x|Y = y) = \prod_{j=i}^{K} P(X^j = x^j|Y = y)$. Oletuksen ollessa voimassa näytteen todennäköisyys kuulua johonkin luokkaan voidaan laskea suhteellisen helposti hyödyntäen Bayesin kaavaa. [3]




Kuva 7: K-lähimmän naapurin luokittelijassa näyte luokitellaan lähellä olevien näytteiden vasteiden perusteella. Mukaillen [3]




Kuva 8: Luokittelevassa tukivektorikoneessa näytteet erotellaan hypertasolla. Mukaillen [16]

*Näyteperusteiset menetelmät* ovat laiskoja oppijoita, joissa ennustaminen tehdään eksplisiittisen mallin opettamisen sijaan vertaamalla uutta näytettä muistissa oleviin näytteisiin. Esimerkki tämän tyypin luokittelijasta on k-lähimmän naapurin luokittelija (*k-Nearest Neighbour classifier*), jossa näytteen luokka määritetään laskemalla, mihin luokkaan näytteen lähellä olevat näytteet kuuluvat. Esimerkki menetelmästä on kuvassa 7.

*Tukivektorikoneet* tekevät hypertasoja (*hyperplanes*), jotka erottelevat dataa lineaarisesti. Hypertaso valitaan niin, että maksimoidaan sen kanssa yhdensuuntaisten marginaalitasojen välinen etäisyys, siten että marginaalitasojen väliin jää vain pieni määrä havaintoja. Usein käytännön sovelluksissa luokat eivät ole tällä tavoin eroteltavissa ja data kuvataan korkeampiulotteiseen avaruuteen (*feature space*), missä se todennäköisemin on mahdollista erotella. Kuvassa 8 on visualisointi menetelmästä. [16]

# 3 Ennakointimallien ennustekyvyn mittarit

Ennustekyvyn (*accuracy*) määrittämiseen on olemassa lukuisia erilaisia mittareita, joita tähän työhön on pyritty keräämään monipuolisesti yleisimmät sisällyttäen. Mittarit kuvaavat eri asioita ennakointimallien ennustekyvystä. Mittareita jatkuvan vasteen malleille on esitelty luvussa 3.1 ja kategorisen vasteen malleille luvussa 3.2.

Osa ennustekyvyn mittareista perustuu residuaaleihin, eli etäisyyteen mallin ennustaman ja havaitun arvon välillä. Kalibrointiin liittyvät mittarit tutkivat havaintojen ja mallin vasteen keskiarvoista yhteneväisyyttä. Hyvin kalibroidulla mallilla ennusteet eivät ole keskimäärin liian suuria tai pieniä. Mallin hyvä kalibrointi ei kuitenkaan vielä tarkoita, että malli osaisi erottaa toisistaan ne näytteet, joille tapahtuu tai ei tapahdu tutkittua lopputulosta. Tähän vaaditaan mallilta erottelukykyä. Tekstissä pyritään erottelemaan residuaaleihin, kalibrointiin ja erotteluun liittyvät mittarit, mutta jaottelu tekeminen ei kuitenkan ole aina täysin yksiselitteistä.

Ennustekyvyn mittareiden käyttö ei ole yksiselitteistä, sillä tulosten tulkintaan ei usein ole käytössä selkeitä kynnysarvoja. Monet mittarit sopivatkin paremmin mallien vertailuun kuin yksittäisen mallin ennustekyvyn arviointiin. Toisaalta mallit, joiden antama kuva datassa olevista yhteyksistä on erilainen, voivat saada samanlaisia tuloksia mittareista, jolloin mallin valinta on vaikeaa [17]. Esimerkiksi, jos käytössä olevassa aineistossa on 20 syötemuuttujaa, joista malliin valitaan neljä, on mahdollisia syötemuuttujien kombinaatioita yli 4000 ja näistä lukuisille mittareiden antamat tulokset ovat samankaltaisia. Tulkintaa vaikeuttaa myös se, että yksittäiset mittarit kertovat vain kapean kuvan mallin toimintakyvystä ja toisaalta voivat antaa keskenään ristiriitaisia tuloksia. Mikään mittareista ei ole täydellinen ja kaikkiin tilanteisiin sopiva, joten niitä käytettäessä tulee olla huolellinen [18]. [19]

Luvussa 2 esitellyn mukaisesti käytettyjä merkintöjä ovat $Y$ vastemuuttujan havaitut arvot, $\hat{Y}$ vastemuuttujan ennustetut arvot, $X$ syötemuuttujan arvot, $N$ otoskoko ja $K$ luokkien määrä. $Y_i^k$ tarkoittaa vastemuuttujan $Y$ havaintoarvoa näytteelle $i$ ja luokalle $k$. Yläindeksi on luonnollisesti tarpeeton jos luokkia ei ole useampia. Merkintä $\bar{Y}$ tarkoittaa Y:n keskiarvoa.

## 3.1 Mittareita jatkuvan vastemuuttujan malleille

Mallien arvioinnissa kannattaa edetä yksinkertaisesta monimutkaisempaan. Havaituista ja ennustetuista vastemuuttujista saa yleiskuvaa vertailemalla niiden keskiarvoja ja variansseja [20]. Regressio-ongelmien tapauksessa monet mitat hyödyntävät residuaaleja, eli havaittujen ja ennustettujen arvojen erotuksia, joita mitoissa summataan yhteen ja skaalataan eri tavoin. Esimerkkejä tällaisistä mitoista ovat keskimääräinen neliövirhe (luku 3.1.1) ja erilaiset korrelaatiokertoimet (luku 3.1.2). Kalibraatiota voi arvioida kalibraatiokuvion (luku 3.1.3) avulla ja erottelua yhteensopivuusindeksillä (luku 3.1.4).

### 3.1.1 Keskimääräinen neliövirhe ja absoluuttinen virhe

Ennusteen ja havaintojen eroa jatkuvilla vastemuuttujilla pyrkivät kuvaamaan keskimääräinen neliövirhe (*mean squared error*, MSE) ja keskimääräinen absoluuttinen

virhe (*mean absolute error*, MAE), jotka voidaan laskea yhtälöillä 1 ja 2. [20]

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \tag{1}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}_i| \tag{2}$$

Usein myös keskimääräisen neliövirheen neliöjuurta (*root mean square error*, RMSE) RMSE $= \sqrt{\text{MSE}}$ käytetään. Nämä mitat summaavat ennusteen ja havaintojen välistä eroa ja antavat sitä kautta kokonaiskuvaa mallin toiminnasta. MAE on näistä vähemmän herkkä poikkeaville havainnoille (*outliers*). Mitoilla on sama yksikkö kuin ennustettavalla muuttujalla, joten arvon tulkinta riippuu datasta. [20]

### 3.1.2 Korrelaatiokertoimet

Pearsonin tulomomentti korrelaatiokerroin (*Pearson product-moment correlation coefficient*) kuvaa muuttujien välistä lineaarista yhteyttä. Se voidaan laskea yhtälöllä 3.

$$r_p = \frac{cov(Y, \hat{Y})}{\sigma_Y \sigma_{\hat{Y}}}, \tag{3}$$

missä $cov(Y, \hat{Y}) = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y}_i)(\hat{Y}_i - \bar{\hat{Y}}_i)$ tarkoittaa kovarianssia ja $\sigma_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y - \bar{Y})^2$ varianssia. Pearsonin tulomomentti korrelaatiokerroin saa arvoja väliltä [-1,1] ja itseisarvoltaan suuri luku tarkoittaa vahvaa lineaarista korrelaatiota. Wilmottin [20] mukaan Pearsonin korrelaatiokerrointa ei tulisi käyttää mallin arviointiin. Sillä ei voi aina saada merkittäviä eroja mallien välille ja toisaalta hyvin erilaisilla havaintojen ja ennusteiden arvoilla voi mitan arvo olla lähellä 1:tä. [21]

Toinen korrelaatioon perustuva mitta on Spearmanin järjestyskorrelaatiokerroin (*Spearman's rank correlation coefficient*), joka voidaan laskea yhtälöllä 4.

$$r_s = \frac{cov(rank(Y), rank(\hat{Y}))}{\sigma_{rankY}\sigma_{rank\hat{Y}}} \tag{4}$$

Yhtälössä $rank(Y_i)$ on $Y_i$ sijaluku, kun $Y$ arvot on järjestetty pienimmästä suurimpaan. Suurin arvo saa sijaluvukseen 1, ja pienin saa $N$. Spearmanin järjestyskorrelaatiokerroin kuvaa lineaarisen yhteyden sijaan kahden muuttujan välistä monotonista riippuvuutta. Myös se saa arvoja väliltä [-1,1]. [22]

Selitysaste (*coefficient of determination*) kuvaa sitä, kuinka suuren osuuden jatkuvan vastemuuttujan havaintoarvojen kokonaisvaihtelusta malli selittää (yhtälö 5). Jos arvo on 1, malli selittää havaintoarvojen vaihtelun täysin, jos taas 0, malli selittää vaihtelua yhtä hyvin kuin datan keskiarvoon asetettu horisontaalinen viiva. Selitysaste yhteys Pearsonin korrelaatiokerroimeen on $R^2 = r_p^2$. [21]

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \tag{5}$$

Selitysasteen $R^2$ arvo kasvaa helposti, kun malliin lisätään syötemuuttujia. Tämän takia on kehitetty korjattu (adjusted) mitta $R^2_{adj}$, joka lasketaan yhtälöllä 6. Sen arvo kasvaa muuttujaa lisätessä vain, jos kasvu on suurempaa kuin sattuman perusteella olisi todennäköistä. Yhtälössä $N_s$ on mallissa olevien syötemuuttujien määrä. [23]

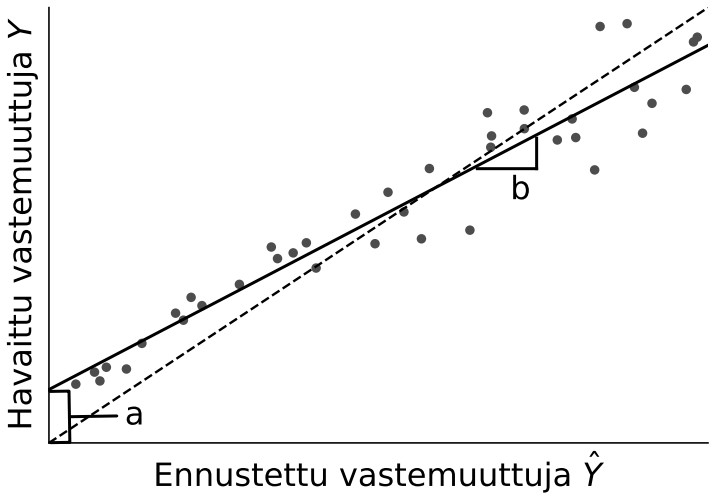
$$R^2_{adj} = 1 - (1 - R^2)\frac{N - 1}{N - N_s - 1} \tag{6}$$

### 3.1.3 Kalibraatiokuvio

Suuren mittakaavan kalibrointi on yksinkertainen mitta, joka määritetään laskemalla ennustettujen ja havaittujen arvojen keskiarvojen erotus, $\bar{\hat{Y}} - \bar{Y} = \frac{1}{N}\sum_i(\hat{Y}_i - Y_i)$. Jos erotus poikkeaa merkittävästi nollasta, ovat arvot systemaattisesti liian suuria tai pieniä. Erotus on yleensä pieni käytettäessä mallin rakentamiseen hyödynnettyä dataa, mutta saattaa erota arvioitaessa muulla aineistolla. [24]

Visuaalisesti regressiomallin kalibraatiota voi tarkastella piirtämällä ennusteet x-akselille ja havainnot y-akselille, mitä kutsutaan kalibraatiokuvioksi. Esimerkki kalibraatiokuviosta regressiomallille on kuvassa 9. Jatkuvan vasteen tapauksessa kyseessä on tavallinen hajontakuvio (*scatter plot*). Tulkinnan helpottamiseksi kuvaan kannattaa piirtää myös 45-asteen suora tasoituskäyrä (*smoothing curve*). [5]

Hyvin kalibroidulle mallille kuvio noudattaa 45-asteen suoraa, eli kalibraatiokuvion kulmakerroin b on lähellä yhtä, eikä kuviossa näy vakiotermiä a. Hyväksyttävä arvo kalibraatiokuvion kulmakertoimelle b riippuu sovelluksesta, mutta esimerkiksi yli 0,9 arvojen on tulkittu olevan riittäviä [25]. Alhainen kulmakerroin kuvastaa usein mallin ylisovittamista ja vakiotermin näkyminen suuren mittakaavan kalibrointivirhettä. Hajontakuviota ja tasoituskäyrää tutkimalla voidaan huomata, että kyseisessä kuviossa on sekä vakiotermiä että sen kulmakerroin on alle 1. [24, 5]




Kuva 9: Kalibraatiokuviossa havaitut vastemuuttujan arvot ovat y-akselilla ja mallin ennustamat x-akselilla. 45-asteen suora (katkoviiva) kuvaa täydellistä kalibraatiota. Yhtenäinen viiva on tasoituskäyrä.

### 3.1.4 Yhteensopivuusindeksi

Erottelevan mallin tulisi systemaattisesti antaa korkeampia tuloksia niille yksilöille, joille havaintoarvo on korkeampi. Tätä voidaan arvioida vertaamalla kaikkia pareja keskenään, määrittämällä toteutuuko sille tämä ehto ja laskemalla kokonaissummat. Yhteensopivuusindeksi (*concordance index*, c) lasketaan yhtälöllä 7. [26]

$$c = \frac{\text{yhteensopiva} + \frac{1}{2}\text{tasatulos}}{\text{kaikki parit}} \tag{7}$$

Tässä pari on yhteensopiva, jos sekä $\hat{Y}_i > \hat{Y}_j$ että $Y_i > Y_j$. Kyseessä on tasatulos, jos joko $\hat{Y}_i = \hat{Y}_j$ tai $Y_i = Y_j$.

Yhteensopivuusindeksi voi saada arvoja väliltä $[0.5, 1]$. Arvo 0.5 vastaa sattumaa. Yleensä yhteensopivuusindeksi saa arvoja väliltä $[0.6, 0.85]$, mutta tämä arvoväli riippuu esimerkiksi tutkittavasta populaatiosta. Arvot $[0.7, 0.8]$ tarkoittavat riittävää ja arvot $[0.8, 0.9]$ erinomaista erottelukykyä. [18]

Yhteensopivuusindeksistä voidaan johtaa Somersin D-indeksi kaavalla 8. Se kuvaa samaa asiaa kuin yhteensopivuusindeksi, mutta se saa arvoja väliltä [-1,1], niin että 0 tarkoittaa ettei erottelua ole.

$$c_D = 2(c - 0.5) \tag{8}$$

Molemmat näistä mitoista ovat helposti tulkittavia. Ne eivät kuitenkaan herkästi erota pieniä erottelukyvyn eroja mallien välillä, sillä ne eivät tunnista eroa esimerkiksi parien (0.01;1) ja (0.9;1) sekä (0.05;0) ja (0.8;1) välillä. [26]

## 3.2 Mittareita kategorisen vastemuuttujan malleille

Myös luokittelumalleja arvioitaessa kannattaa aloittaa yksinkertaisilla havainnollistuksilla kuten vertailemalla eri luokkien ennustettujen ja havaittujen näytteiden määriä esimerkiksi sekaannusmatriisin avulla (luku 3.2.6). Tärkeä peruskäsite on luokittelun tarkkuus (luku 3.2.1). Kalibraation ja erottelun mittaaminen on tärkeää myös luokittelun tapauksessa, joista kumpaankin on kehitetty lukuisia mittoja.

Moniluokkaisissa ja moninimikkeisissä tapauksissa mitat saattavat olla binäärisistä mitoista johdettuja mikro- tai makrokeskiarvoja (*mikro-averaging*, *makro-averaging*). Makro-keskiarvoissa lasketaan ensin luokkakohtaiset arvot, jotka lopuksi yhdistetään. Mikro-keskiarvoissa puolestaan lasketaan ensin summia, joista lasketaan mitta. Makro-keskiarvoistus kohtelee kaikkia luokkia samalla tavoin ja mikro-keskiarvoistus suosii isoja luokkia. Moninimike luokittelussa ennustekyvyn laskeminen on usein haastavampaa, eikä siihen perehdytä syvällisesti tässä työssä. Aiheesta ovat kirjoittaneet esimerkiksi Tsoumakas ym. [27] ja Zhang ym. [12]. Yhteenveto tässä työssä esitellyistä mittareista erilaisille luokittelumalleille löytyy taulukosta 4. [14]

### 3.2.1 Luokittelun tarkkuus

Luokittelun tarkkuus (*classification accuracy*) on oikein luokiteltujen havaintojen prosenttiosuus kaikista havainnoista (yhtälö 9). Virhe-aste (*error rate*) on vastaavasti

Taulukko 4: Mittareita kategorisen vasteen malleille

| | Binäärinen | Moniluokkainen | Moninimikkeinen |
|---|---|---|---|
| Yleinen | ACC<br>Sekaannusmatriisi<br>Brierin pistemäärä | ACC<br>Sekaannusmatriisi<br>Brierin pistemäärä | ACC, $\text{ACC}_{ML}$<br>Hammingin etäisyys |
| Kalibrointi | Hosmer-Lemeshow mitta<br>Suuren mittakaavan kalibrointi<br>Kalibraatiokuvio<br>Validaatiokuvio | Suuren mittakaavan kalibrointi | |
| Erottelu | Sensitiivisyys<br>Spesifisyys<br>AUC<br>Yhteensopivuusindeksi<br>Erottelun kulmakerroin<br>RS | Sensitiivisyys<br>Spesifisyys<br>$\text{AUC}_M$<br>Yhteensopivuusindeksi<br>RS | $\text{Sensitiivisyys}_{ML}$<br>$\text{Spesifisyys}_{ML}$<br>$\text{AUC}_{ML}$ |
| Uuden muuttujan lisäys | | Uudelleenluokittelutaulu<br>NRI, IDI | |

virheellisesti luokiteltujen havaintojen osuus kaikista havainnoista, eli 1-ACC. Mitat ovat yksinkertaisia ja laajasti käytössä ja ne ovat monesti pohjana monimutkaisemmille mittareille. Niiden antama informaatio on kutienkin rajallista ja esimerkiksi otosjoukon epätasapaino vaikuttaa niihin vahvasti. Mittoja tulkittaessa tuleekin olla huolellinen. [16]

$$ACC = \frac{1}{N} \sum [\hat{Y}_i = Y_i] \tag{9}$$

Moninimikeluokittelussa yhtälön 9 mukaista mittaa $ACC$ voidaan kutsua myös osajoukon tarkkuudeksi (*subset accuracy*). Erityisesti nimikejoukon ollessa iso se on hyvinkin tiukka vaatiessaan kaikkien näytteelle ennustettujen nimikkeiden olevan oikein. Toinen tapa laskea luokittelun tarkkuus on jakaa havaittujen ja mallin ennustamien nimikkeiden joukkojen leikkaukseen kuuluvien nimikkeiden lukumäärä niiden yhdisteeseen kuuluvien lukumäärällä (yhtälö 10). [12]

$$ACC_{exam} = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|} \tag{10}$$

### 3.2.2 Brierin pistemäärä

Brierin pistemäärällä (*Brier score*) voidaan mitata havaintoarvojen ja mallin antamien tulosten yhteyttä kategorisille vastemuuttujille. Se voidaan laskea yhtälöllä 11, missä merkintä $h^j(X_i)$ tarkoittaa mallin antamaa todennäköisyyttä sille, että näyte $i$ kuuluu luokkaan $j$. Brierin pistemäärä saa arvon 0, jos kaikki ennusteet ovat osuneet oikeaan. [28, 29]

$$BS = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}(h^j(X_i) - Y_i^j)^2 \tag{11}$$

### 3.2.3 Hammingin etäisyys

Hammingin etäisyys (*Hamming loss*) laskee väärinluokiteltujen näyte-nimikeparien osuutta. Se voidaan laskea yhtälöllä 12. [12]

$$hamming = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}|\hat{Y}_i \Delta Y_i| \tag{12}$$

Yhtälössä $\Delta$ on symmetrinen erotus joukkojen välillä. Se huomioi ne nimikeparit, joiden ennustettu ja havaittu tulos eroavat. Niissä tapauksissa joko ennustettua nimikettä ei ole havaittu tai havaittua nimikettä ei ole ennustettu.

### 3.2.4 Hosmer-Lemeshown mitta

Hosmer-Lemeshown mitta (*Hosmer-Lemeshow measure*) tarkastelee binäärisen vastemuuttujan mallin kalibraatiota jakamalla tulokset osajoukkoihin. Mitan laskeminen koostuu seuraavista vaiheista [30]:

1. Järjestetään näytteet ennustetun todennäköisyyden $h(X_i)$ perusteella suurimmasta pienimpään.

2. Jaetaan järjestetyt havainnot G samankokoiseen luokkaan. Usein $G = 10$, mutta se voi olla muutakin erityisesti suurella otoskoolla.

3. Lasketaan kullekin luokalle g seuraavat mitat summaamalla luokkaan kuuluvien näytteiden määrät yhteen: $N_{O0,g}$: havainnoidut määrät vasteelle 0, $N_{E0,g}$: ennustetut määrät vasteelle 0, $N_{O1,g}$: havainnoidut määrät vasteelle 1, $N_{E1,g}$: ennustetut määrät vasteelle 1.

4. Lasketaan Hosmer-Lemeshown mitta $\hat{C}_G$ (yhtälö 13). Mitta noudattaa $\chi^2$-jakaumaa vapausasteilla $v = (G - 2)$.

$$\hat{C}_G = \sum_{g=1}^{G}[\frac{(N_{O0,g} - N_{E0,g})^2}{N_{E0,g}} + \frac{(N_{O1,g} - N_{E1,g})^2}{N_{E1,g}}] \tag{13}$$

5. Lasketaan saadulle mitalle p-arvo $\chi^2$ -testillä [21]. Nollahypoteesina on, että havaitut ja odotetut todennäköisyydet ovat yhtenevät kaikissa ryhmissä. Vaihtoehtoisena hypoteesina on, että todennäköisyydet eivät ole yhtenevät.

Kuten muidenkin $\chi^2$ testien tapauksessa, otoskoon kasvaessa Hosmer-Lemeshown mitan todennäköisyys hylätä huonosti sopiva malli kasvaa. Ennakointimallia testattaessa tämä ei ole toivottu ominaisuus, sillä hyvän testin todennäköisyys hylätä malli pitäisi olla riippumaton käytetystä otoskoosta eikä oletuksena ole rakentaa täysin virheetöntä mallia. Tätä ratkaisemaan on kehitetty menetelmiä kuten luokkien määrän vaihtelu eri otoskoolla. Mittaa on kritisoitu myös, koska sen antama tulos riippuu valitusta ryhmittelystä, mitan voimakkuus on alhainen pienillä otoskoilla ja tuloksena on tulkinnanvaraa jättävä p-arvo [24]. Mitta ei myöskään kerro mihin suuntaan mallissa on virhekalibrointia [31]. [30]

### 3.2.5 Kalibraatiokuvio

Myös luokittelumalleille kalibraatiokuvio (luku 3.1.3) on hyödyllinen työkalu. Suuren mittakaavan kalibroinnin voi moniluokkaisen ja moninimikkeisen vastemuuttujan tapauksessa laskea kullekin luokalle erikseen. Jokaiselle luokalle lasketaan havaittu ja ennustettu osuus näytteistä, jotka kuuluvat kyseiseen luokkaan ja otetaan niistä erotus. [24]

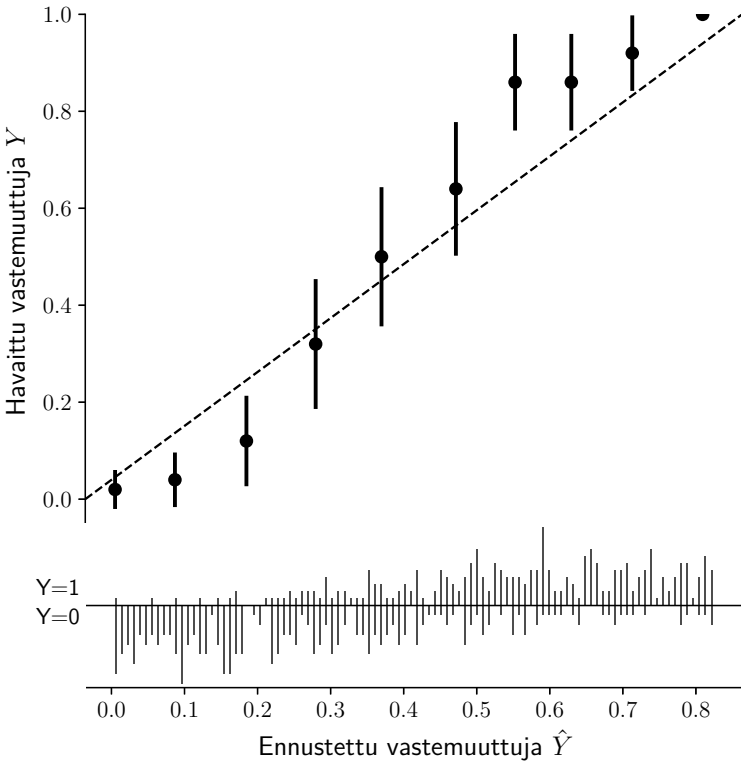
Binääriselle vasteelle voi piirtää kalibraatiokuvion samalla tavalla kuin jatkuvalle, mutta Y-akseli saa vain arvoja 0 tai 1. Tulkintaa helpottamaan kannattaa piirtää tasoituskäyrä. Jos malli antaa binääriselle vasteelle luokittelun 0/1 sijaan todennäköisyyden saada arvo yksi (eli luvussa 2.2 määritellyn funktion $h(x)$), voidaan näytteet ryhmitellä luokkiin $h(x)$ mukaan. Tämän jälkeen eri luokille voidaan laskea havaittujen ja ennustettujen arvojen keskiarvot 95% luottamusväleineen ja piirtää niistä hajontakuvio luokittain. Kuvion voi tulkita olevan Hosmer-Lemeshown mitan 3.2.4 visualisointi. Tämän tyyppinen analyysi tuottaa helposti tulokseksi hyvän kalibraation, sillä keskiarvoistaminen vähentää tulosten satunnaisuuden merkitystä [18]. Verrattuna Hosmer-Lemeshown mittaan kalibraatiokuvio on informatiivisempi, sillä p-arvon lisäksi siitä saa käsitystä vaikutuksen suuruudesta ja luottamusväleistä. Lisäksi havaintoja ei tarvitse jakaa luokkiin. [24, 5]

Kalibraatiokuvion voi binäärisen vasteen tapauksessa laajentaa validaatiokuvioksi (*validation plot*). Tällöin ennusteiden jakauma ilman vastetta ja sen kanssa piirretään kuvaajan alalaitaan. Kalibraation lisäksi validaatiokuviosta voi tulkita erottelua tutkimalla havaittujen vasteiden jakaumaa eri ennusteiden arvoväleillä. Lisäksi mallin hyödyllisyyttä voidaan määrittää tutkimalla, kuinka suuri osa ennusteita ylittää tai alittaa jonkin relevantin kynnysarvon. [5]

Kuvassa 10 on esimerkki validaatiokuvaajasta binäärisen vasteen mallille. Yläosassa on piirretty ennusteen mukaan desiileihin luokitelluille näytteille lasketut keskiarvot ennusteelle ja havainnoille. Havaintojen keskiarvolle on myös piirretty luottamusvälit. Kuvaajan ala-osassa puolestaan on pylväsdiagrammi ennusteiden frekvenssistä kuvattuna erikseen näytteille, joilla Y=1 (ylempi) ja Y=0 (alempi).

### 3.2.6 Sekaannusmatriisi, sensitiivisyys ja spesifisyys

Sekaannusmatriisilla (*confusion matrix*) kuvataan sitä, miten näytteiden lukumäärät jakautuvat ennustetun ja havaitun luokan mukaan [3, luku 9]. Esimerkki sekaannusmatriisista kolmen luokan tapauksessa on taulukossa 5. Merkintä $N_{A,B}$ tarkoittaa




Kuva 10: Validaatiokuvio piirrettynä luokittelemalla näytteet ennusteen arvon mukaan yhtä suuriin luokkiin. Y-akselilla on havaittujen arvojen keskiarvo ja x-akselilla ennusteen keskiarvo. Lisäksi on piirretty havaintojen keskiarvon 95% luottamusväli. 45-asteen suora (katkoviiva) kuvaa täydellistä kalibraatiota. Alareunan kuvaajassa on lisäksi kuvattu ennusteiden frekvenssit erikseen tilanteille, joissa havainto on 0 ja 1. Mukailtu [5]

niiden näytteiden lukumäärää, joiden ennustettu luokka on $A$ ja havaittu luokka on $B$. Diagonaalilla olevissa soluissa ovat ne näytteet, jotka on luokiteltu oikein (kuvassa tummennettuna). Sekaannusmatriisin mukaista jaottelua käytetään pohjana moniin mittoihin.

Taulukko 5: Sekaannusmatriisiin kirjataan ennustetut ja havaitut näytteiden lukumäärät eri luokille. [3]

| Ennustettu luokka | Havaittu luokka | | |
|---|---|---|---|
| | Luokka A | Luokka B | Luokka C |
| Luokka A | $N_{A,A}$ | $N_{A,B}$ | $N_{A,C}$ |
| Luokka B | $N_{B,A}$ | $N_{B,B}$ | $N_{B,C}$ |
| Luokka C | $N_{C,A}$ | $N_{C,B}$ | $N_{C,C}$ |

Keskeisiä erotteluun liittyviä käsitteitä kategorisen vastemuuttujan tapauksessa ovat sensitiivisyys ja spesifisyys. Sensitiivisyys (*sensitivity*, *recall*) tarkoittaa mallin

kykyä luokitella oikein ne näytteet, joille tutkittava ehto täyttyy. Sitä voi kutsua myös oikeiden positiivisten osuudeksi. Spesifisyys (*specificity*, *1-precision*) puolestaan tarkoittaa mallin kykyä luokitella oikein ne näytteet, joille tutkittava ehto ei täyty. Spesifisyys on myös 1-väärien positiivisten osuus. Usein käytettyjä termejä binäärisen vasteen malleille ovat positiivinen ennustearvo (*positive predictive value*, PPV) sekä negatiivinen ennustearvo (*negative preditive value*, NPV). PPV tarkoittaa todennäköisyyttä sille, että havainto on 1 ennusteen ollessa 1 ja NPV todennäköisyyttä, että havainto on 0 ennusteen ollessa 0. PPV ja NPV arvot riippuvat siitä, kuinka todennäköistä populaatiossa on saada arvo 0 tai 1. [32]

Termejä voi havainnollistaa taulukolla 6, jossa sekaannusmatriisi on kuvattu luokan $j$ näkökulmasta. Tällöin toteutumisvaihtoehtoja on neljä: oikea positiivinen (*true positive*, TP), väärä positiivinen (*false postitive*, FP) väärä negatiivinen (*false negative*, FN) ja oikea negatiivinen (*true negative*, TN). Taulukossa esimerkiksi $N_{j,j^\complement}$ tarkoittaa niiden näytteiden määrää, joille ennustettu luokka on $j$ ja havaittu luokka on jokin muu kuin $j$, joten kyseessä on väärien positiivisten määrä $N_{FP}$. Merkinnät ovat vastaavat myös muille vaihtoehdoille. Luokittelu voi käytännössä olla esimerkiksi, että luokka $j$ tarkoittaa sairautta ja ei-$j$ sitä, että sairautta ei ole. Tällöin $PPV = \frac{N_{TP}}{N_{TP}+N_{FP}}$ ja $NPV = \frac{N_{TN}}{N_{TN}+N_{FN}}$.

Taulukko 6: Sekaannusmatriisi luokan $j$ näkökulmasta.

| Ennustettu luokka | Havaittu luokka | |
|---|---|---|
| | $j$ | Ei-$j$ |
| $j$ | $N_{j,j} = N_{TP}$ | $N_{j,j^\complement} = N_{FP}$ |
| Ei-j | $N_{j^\complement,j} = N_{FN}$ | $N_{j^\complement,j^\complement} = N_{TN}$ |

Näillä merkinnöillä sensitiivisyyden voi laskea yhtälöllä 14 ja spesifisyyden yhtälöllä 15 binäärisille tai moniluokkaisille vastemuuttujille. Näistä voi laskea myös todennäköisyyssuhteen (*likelihood ratio*, LR) yhtälöllä 16. LR kuvaa sitä, kuinka paljon todennäköisempää on, että luokkaan $j$ ennustettu näyte myös havaintojen puolesta kuuluu luokkaan $j$ kuin sellainen, jolle ennuste on ei-$j$. [14]

$$\text{Sensitiivisyys} = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{14}$$

$$\text{Spesifisyys} = \frac{N_{TN}}{N_{TN} + N_{FP}} \tag{15}$$

$$\text{LR} = \frac{\text{Sensitiivisyys}}{\text{1-Spesifisyys}} \tag{16}$$

Moninimikkeisessä tapauksessa mitat voidaan laskea vastaavasti, mutta toki yleistettynä joukoille. Tällöin sensitiivisyys voidaan laskea kaavalla 17 ja spesifisyys 18. [12]

$$\text{Sensitiivisyys}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i|}{|\hat{Y}_i|} \tag{17}$$

$$\text{Spesifisyys}_{ML} = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i|}{|Y_i|} \tag{18}$$

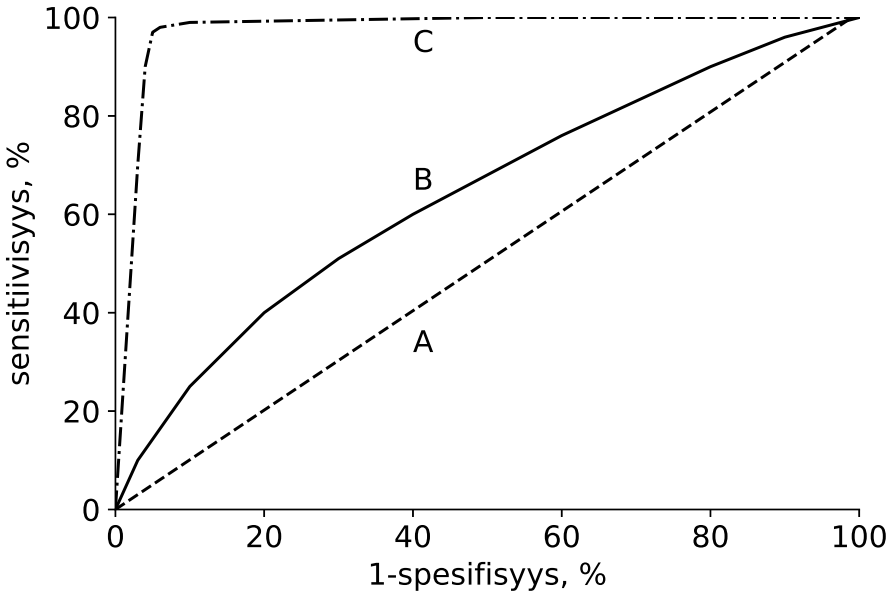
Testin korkeasta sensitiivisyydestä on hyötyä silloin, kun väärän negatiivisen tuloksen kustannus on suuri. Vastaavasti korkea spesifisyys on tärkeää, jos väärän positiivisen tuloksen kustannus on suuri. Testin sensitiivisyyteen ja spesifisyyteen vaikuttaa kynnysarvo, jonka perusteella testin tulos määritellään positiiviseksi tai negatiiviseksi. Yleisesti, jos kynnysarvoa nostetaan, on vähemmän vääriä positiivia mutta enemmän vääriä negatiivisia. Mallin spesifisyys siis kasvaa ja sensitiivisyys pienenee. Vastaavasti kynnysarvoa laskettaessa mallin spesifisyys pienenee ja sensitiivisyys kasvaa. [32]

### 3.2.7 ROC-käyrä ja sen alle jäävä pinta-ala

ROC-käyräksi kutsutaan eri kynnysarvoille laskettua kuvaajaa, jossa x-akselilla on mallin 1-spesifisyys ja y-akselilla on sensitiivisyys [32]. Esimerkki ROC-käyristä on esitelty kuvassa 11. ROC-käyrän alle jäävä pinta-ala (AUC) kuvaa mallin erottelukykyä. Sen voidaan tulkita tarkoittavan todennäköisyyttä sille, että malli luokittelee oikein sattumanvaraisesti valitut ehdon täyttävän ja ei täyttävän havainnon [33]. Jos valitaan kaksi näytettä, joista toisen havainto on 0 ja toisen 1, antaa AUC todennäköisyyden sille, että mallin ennuste on korkeampi näytteelle havaintoarvolla 1 [34]. Jos testin AUC arvo on lähellä yhtä, on testillä sekä korkea sensitiivisyys että spesifisyys eli sen erottelukyky on hyvä. Jos taas AUC on lähellä 0.5, mallin erottelukyky on huono. AUC arvon mittaamisen etuja on esimerkiksi se, että se ei ole riippuvainen valitusta luokkiin erottelevasta kynnysarvosta eikä vastemuuttujan jakaumasta. [35]

Yleisesti AUC arvot-[0.9,1] tarkoittavat erinomaista, [0.8,0.9] hyvää, [0.7,0.8] kohtuullista ja [0.6,0.7] heikkoa erottelukykyä. Erottelua voidaan pitää hyödyllisenä, jos AUC-arvo on yli 0.75. Tämä arvo on yleinen kirjallisuudessa, mutta myös muita arvoja kuten 0.8 on esitetty [36]. Jossain tilanteissa on järkevää laskea AUC vain tietystä osasta ROC-käyrää. Esimerkiksi jos olemme kiinnostuneita vääristä positiivisista mutta emme vääristä negatiivisista. [25]

AUC yleistyy myös moniluokkaiseen ja moninimikkeiseen luokitteluun, jos malli antaa todennäköisyyden, että näyte $i$ kuuluu luokkaan $j$ eli arvon $h^j(X_i)$. Tarkastellaan tilannetta ensin binäärisen luokittelun (luokat 0 ja 1) tapauksessa. Määritetään $g_i = h^0(X_i)|0, i = 1, ..., N_0$, joka tarkoittaa $i$:n näytteen ennustettua todennäköisyyttä kuulua luokkaan 0. Joukkoon otetaan todennäköisyydet vain niille näytteille, joille havainto on 0. Niiden lukumäärä on $N_0$. Vastaavasti määritetään $e_i = h^1(X_i)|1, i = 1, ..., N_1$. Yhdistetään joukot $g_i$ ja $e_i$ ja järjestetään ne kasvavaan järjestykseen. Luku $r_i$ on $i$:nnen luokkaan 0 kuuluvan näytteen sijaluku. Tällöin niiden yhdestä luokan 0 ja yhdestä luokan 1 havainnoista koostuvien parien lukumäärä, joille luokan 1 havainnolla on ennusteen mukaan pienempi todennäköisyys kuulua luokkaan 0 on $\sum_{i=0}^{N_0}(r_i - i) = S_0 - \frac{N_0(N_0+1)}{2}$. Tässä $S_0 = \sum_{i=0}^{N_0} r_i$. Yhteensä mahdollisia pareja on $N_0N_1$ kappaletta, joten todennäköisyys, että sattumanvaraisesti valitulla luokkaan 1 kuuluvalla havainnolla on ennusteen mukaan pienempi todennäköisyys




Kuva 11: Esimerkkejä ROC-käyristä. A: ei erottelua (AUC=0.5), B: tyypillinen tulos (AUC=0.5-1.0) ja C: täydellinen erottelu (AUC=1.0). Mukailtu [32]

kuulua luokkaan 0 kuin sattumanvaraisesti valitulla luokkaan 0 kuuluvalla havainnollam, voidaan laskea yhtälöllä 19. Yhtälöllä 19 saadaan estimaatti AUC-mitalle, mutta AUC lasketaan yleensä integroimalla ROC-käyrän alle jäävä pinta-ala. [34]

$$AUC = S_0 - \frac{N_0(N_0+1)}{2})/N_0N_1 \tag{19}$$

Oletetaan sitten, että luokkia $K > 2$. Luokat on numeroitu $0, 1, ..., K-1$, eikä niiden keskinäisellä järjestyksellä ole väliä. Luokista muodostettuja pareja on yhteensä $K(K-1)$ kappaletta. Tällöin todennäköisyys voidaan summata luokkien yli ja saadaan erottelun mitta $AUC_M$ (yhtälö 20).

$$AUC_M = \frac{2}{K(K-1)} \sum_{i<j} \frac{S_i - \frac{N_i(N_i+1)}{2}}{N_iN_j} \tag{20}$$

AUC voidaan laskea myös moninimikemalleille yhtälöllä 21. [12]

$$AUC_{ML} = \frac{1}{K}\sum_{j=1}^{K} AUC_j = \frac{1}{K}\sum_{j=1}^{K} \frac{|\{(X', X'')|h^j(X') \geq h^j(X''), (X', X'') \in Z_j \times \bar{Z}_j\}|}{|Z_j||\bar{Z}_j|} \tag{21}$$

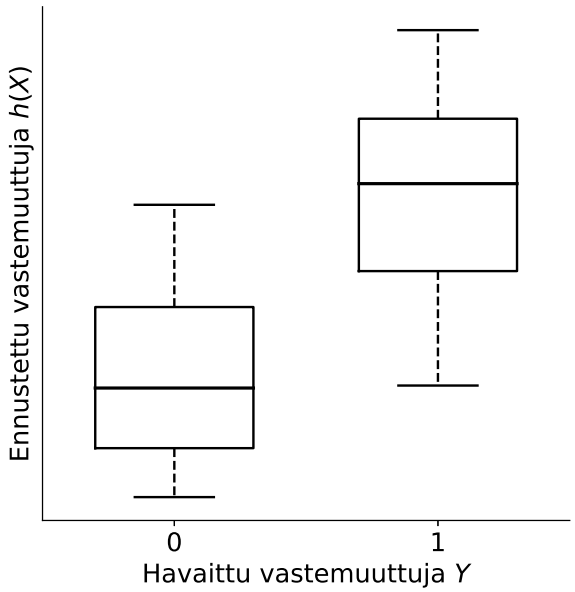
missä $Z_j = \{X_i|y_j \in Y_i, i = 1, .., N\}$ ($\bar{Z}_j = \{X_i|y_j \notin Y_i, i = 1, .., N\}$) tarkoittaa otosjoukkoa, johon kuuluu (ei kuulu) nimikettä $y_i$.

### 3.2.8 Muita erottelumittareita

Binääriselle ja moniluokkaiselle vasteelle yhteensopivuusindeksi voidaan laskea vastaavasti kuin jatkuvalle vastemuuttujalle. Se on esitelty luvussa 3.1.4. Jos vastemuuttuja on binäärinen, yhteensopivuusindeksi on sama kuin AUC [37].

Erottelun kulmakerroin (*discrimination slope*) on yksinkertainen tapa arvioida erottelua binäärisen vasteen tapauksessa. Se lasketaan määrittämällä ennusteiden keskiarvo näytteille, joiden havaintoarvo on 1 ($\hat{Y}_{i,1}$) ja vähentämällä siitä ennusteiden keskiarvo niiden näytteiden tapauksessa, joiden havainto on 0 ($\hat{Y}_{i,0}$). Tämä on esitetty yhtälössä 22. Mittaria voi visualisoida piirtämällä ennusteista laatikkokuvion (*box plot*) erikseen näytteille, joiden havaintoarvo on 0 ja 1 (kuva 12). Kuvioissa näkyy vähemmän päällekkäisyyksiä paremmin erottelevalle mallille. [5]

$$DS = \frac{\sum_i \hat{Y}_{i,1}}{N_1} - \frac{\sum_i \hat{Y}_{i,0}}{N_0} \tag{22}$$




Kuva 12: Erottelun kulmakerrointa voi havainnollistaa laatikkokuviolla, jotka piirretään erikseen vasteille 0 ja 1. Kuvan tilanteessa erottelun kulmakertoimen arvo on 0.3. Mukailtu [5]

Hyvä riskimalli antaa arvoja, joiden pohjalta voi päättää toimenpiteistä. Se voi tarkoittaa esimerkiksi sitä, että malli luokittelee valtaosan näytteistä joko suurimman tai pienimmän riskin mukaisiin luokkiin, jolloin tulosten tulkinta on helppoa. Toisaalta, jos malli ennustaisi kaikille lähes samansuuruista riskiä, olisi näytteiden välille vaikeaa saada eroja. Riskin osituskyky (*risk stratification capacity*) tarkoittaa riskiluokkia ennustavan mallin kykyä jakaa näytteet relevantteihin riskiluokkiin. Se, mitä relevantti riskiluokka tarkoittaa, riippuu siitä, mitä pyritään ennustamaan. Riskin osituskyky lasketaan kaavalla 23. [38]

$$RS = \frac{N_{\text{relevantit luokat}}}{N} \tag{23}$$

### 3.2.9 Uudelleenluokittelu

Malleja kehitettäessä on joskus mahdollista lisätä malliin syötemuuttujia. Haasteena tässä on kuitenkin usein se, että uudet muuttujat saattavat kasvattaa monia ennustemallien mittareita vain vähän vaikka niiden tilastollinen yhteys ennustettavaan muuttujaan olisi suuri. Muuttujien merkitystä voi siksi olla vaikeaa arvioida. Riskimalleille muuttujien lisäämisen paremmin huomioon ottavia tilastollisia mittareita ovat NRI (*net reclassification improvement*) ja IDI (*integrated discrimination index*). Tilannetta voi havainnollistaa uudelleenluokittelutaulukolla (*reclassification table*), josta löytyy esimerkki taulukosta 7 [39]

Taulukko 7: Esimerkki uudelleenluokittelutaulukosta. Tauluun kirjataan luokkiin kuuluvien näytteiden määrät mallille uuden muuttujan kanssa ja ilman sitä. Taulussa erotellaan näytteet, joille realisoituu ja ei realisoidu, jonka riskiä ennustetaan. [39]

| | **Malli uudella muuttujalla** | | |
|---|---|---|---|
| **Malli ilman uutta muuttujaa** | Korkea riski | Matala riski | Summa |
| **Näytteet, joille riski realisoituu** | | | |
| Korkea riski | 20 | 5 | 25 |
| Matala riski | 10 | 35 | 45 |
| Summa | 30 | 40 | 70 |
| **Näytteet, joille riski ei realisoidu** | | | |
| Korkea riski | 20 | 5 | 25 |
| Matala riski | 250 | 70 | 320 |
| Luokka B | 45 | 200 | 245 |
| Summa | 295 | 270 | 565 |

Uudelleenluokittelutaulukossa havainnollistetaan sitä, miten muuttujan lisääminen malliin vaikuttaa näytteiden sijoittumiseen eri luokkiin. Tauluun kirjataan kuhunkin luokkaan kuuluvien näytteiden määrät alkuperäisellä ja laajennetulla mallilla. Pencina ym. [39] erottelevat uudelleenluokittelutaulukossa ne näytteet, joille riski realisoituu ja ei realisoidu, eli joko tapahtuu tai ei tapahdu asiaa, jonka riskiä ennustetaan. Syynä on se, että olennaista mallin toimintakyvyn paranemisen kannalta on, että siirto riskiluokissa tapahtuu oikeaan suuntaan. Jos näytteitä joille riski realisoituu liikkuu kategorioissa "ylöspäin", eli kohti suurempaa riskiä, luokittelu paranee. Jos taas "alaspäin", luokittelu huononee. Tulkinta on päinvastaista näytteille, joille riski ei realisoidu. [39]

Määritetään funktio $v(i)$ kuvaamaan näytteiden liikkumista luokkien välillä (yhtälö 24). Uudelleenluokittelua voidaan havainnollistaa summaamalla $v(i)$ niille näytteille, joille riski realisoituu ja vähentämällä tästä summa niille, joille riski ei realisoidu (yhtälö 25). Näin summataan yhteen ne osuudet, jotka liikkuvat riskiluokissa oikeaan

suuntaan. [39]

$$v(i) = \begin{cases} 1, & \text{jos yksilö i liikkuu kategorioissa ylöspäin} \\ 0, & \text{jos yksilö i pysyy samassa kategoriassa} \\ -1, & \text{jos yksilö i liikkuu kategorioissa alaspäin} \end{cases} \tag{24}$$

$$NRI_1 = \frac{\sum_{\text{i, realisoituu}} v(i)}{N_{\text{realisoituu}}} - \frac{\sum_{j,\text{ei realisoidu}} v(j)}{N_{\text{ei realisoidu}}} \tag{25}$$

NRI:n arvo riippuu valituista kategorioista. Se voidaan välttää tekemällä luokittelusta niin tarkkaa, että jokainen näyte on omassa luokassaan. Diskreetin $v(i)$ sijasta lasketaan erotus ennustetuissa todennäköisyyksissä vanhan ja uuden mallin välillä. Tällöin NRI otokselle voidaan laskea kaavalla 26. [39]

$$NRI_2 = (\hat{p}_{\text{ylös, realisoituu}} - \hat{p}_{\text{alas, realisoituu}}) - (\hat{p}_{\text{ylös, ei realisoidu}} - \hat{p}_{\text{alas, ei realisoidu}}) \tag{26}$$

Yhtälössä $\hat{p}$ ovat otoksesta laskettuja estimaatteja todennäköisyyksille, esimerkiksi

$$\hat{p}_{\text{ylös, realisoituu}} = \frac{N_{\text{ylös, realisoituu}}}{N_{\text{realisoituu}}}$$

Toinen [39] esittelemä mitta, IDI, voidaan laskea yhtälöllä 27. Sen voidaan tulkita olevan erotus keskimääräisen sensitiivisyyden ja 1-spesifisyyden paranemisen välillä. $IDI$ on yhteydessä muihin mittoihin, sillä se on myös uuden ja vanhan mallin erottelun kulmakertoimien, Pearsonin korrelaatiokertoimien tai skaalattujen Brierin pistemäärien erotus [5]. [39]

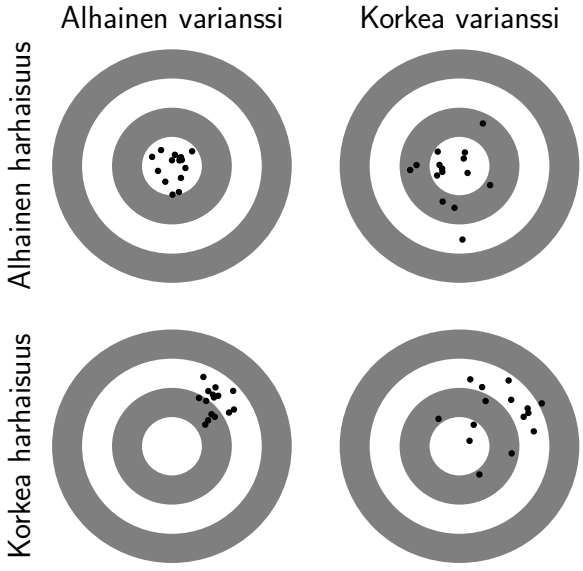
$$IDI = (\bar{\hat{p}}_{\text{uusi, realisoituu}} - \bar{\hat{p}}_{\text{uusi, ei realisoidu}}) - (\bar{\hat{p}}_{\text{vanha, realisoituu}} - \bar{\hat{p}}_{\text{vanha, ei realisoidu}}) \tag{27}$$

Yhtälössä $\bar{\hat{p}}$ tarkoittavat mallin antamien todennäköisyyksien keskiarvoja. Esimerkiksi $\bar{\hat{p}}_{\text{uusi, realisoituu}}$ tarkoittaa uudella mallilla laskettujen todennäköisyyksien keskiarvoa niille näytteille, joille riski realisoituu.

# 4 Ennakointimallien yleistyvyyden tutkiminen

Luvussa 3 esiteltiin ennustekyvyn mittareita. Jos mittareiden arvoja lasketaan samoilla havainnoilla kuin malli rakennetaan, voidaan määrittää, kuinka hyvin malli sopii kyseiseen dataan. Mallin näennäinen ennustekyky tässä joukossa on optimistisempi kuin se olisi muussa edes samasta populaatiosta otetussa joukossa. Mallin tulisi kuitenkin toimia myös uudella aineistolla todellisessa käyttöympäristössä, minkä vuoksi yleistyvyyden tutkiminen on tärkeää mallin suorituskykyä arvioitaessa. Yleistyvyys voidaan jakaa toistettavuuteen (*reproducibility*) ja siirrettävyyteen (*transportability*). Toistettavuus kuvaa mallin ennustekykyä samasta populaatiosta otetuille, mutta ei mallin rakentamiseen käytetyille näytteille. Siirrettävyyden tapauksessa kyse on puolestaan toisen, mutta keskeisiltä ominaisuuksiltaan samankaltaisen populaation näytteistä.

Yleistyvyyttä arvioidaan usein otoksella, joka on otettu samasta populaatiosta kuin mallin kehittämiseen käytetty joukko. Tätä kutsutaan sisäiseksi validoinniksi (*internal validation*) ja voidaan tutkia toistettavuutta. Sisäisen validoinnin toteuttamista käsitellään luvussa 4.1 ja datan jakamista mallin opettamista ja arviointia varten luvussa 4.2. Sisäisen validoinnin lisäksi voidaan tehdä ulkoista validointia (*external validation*) eri populaatiosta olevalla aineistolla. Tällöin voidaan tutkia myös mallin siirrettävyyttä, mistä on lisää luvussa 4.3. Riittävä otoskoko on keskiössä hyvän mallin rakentamisessa. Lisäämällä mallin opettamiseen käytettävää data, voidaan mallin ennustekykyä parantaa tiettyyn rajaan asti. Tästä käyttäytymisestä ja riittävän otoskoon määrittämisestä kerrotaan luvussa 4.4. [29]



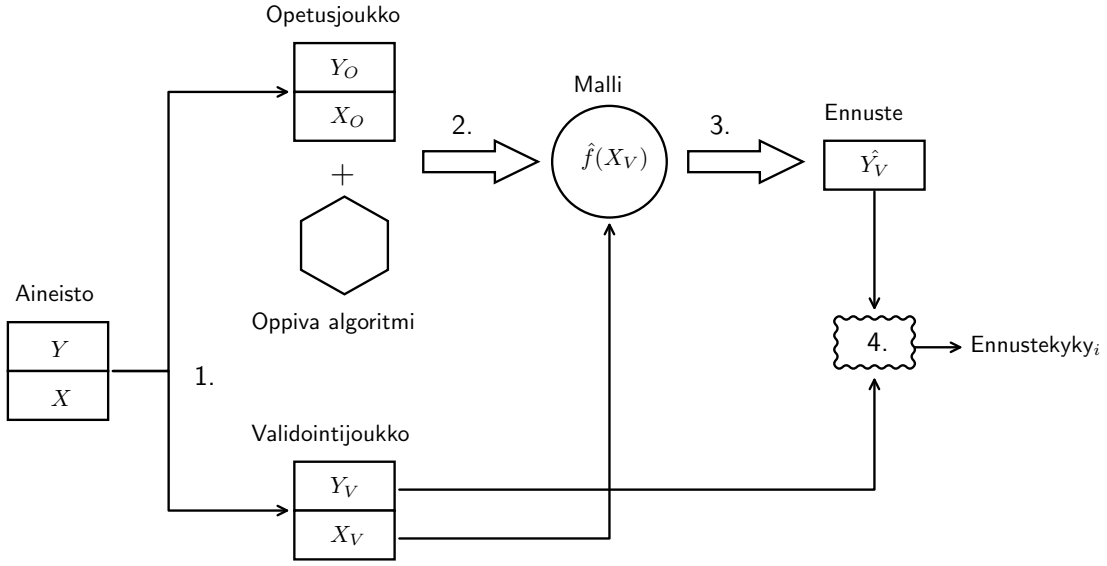

Kuva 13: Harhaisuus kuvaa estimaatin odotusarvon ja oikean arvon yhteneväisyyttä ja varianssi estimaatin satunnaista vaihtelua. [40]

Mallia arvioitaessa muodostetaan estimaatteja $\hat{\theta}$ ennustekyvyn mittareista $\theta$. Mittareita voidaan vain estimoida, sillä otoksen taustalla oleva jakauma ei ole täysin tiedossa ja otoskoko on rajallinen. Esimerkki tällaisesta estimaatista on

luokittelun tarkkuus. Estimaatteja voidaan arvioida niiden harhaisuuden ja varianssin näkökulmasta. Harhaisuus $Bias(\theta) = E(\hat{\theta}) - \theta$ kuvaa sitä, kuinka hyvin estimaatin odotusarvo vastaa mittarin oikeaa arvoa. Jos $Bias = 0$, on estimaatti harhaton. Jos estimaatti antaa oikeaa arvoa huonompia tuloksia, sitä kutsutaan pessimistiseksi, ja jos taas parempia, se on optimistinen. Estimaatin varianssi $Var(\theta) = E[(\hat{\theta} - E\hat{\theta})^2]$ kuvaa sitä, kuinka paljon estimaatti vaihtelee, jos opetusprosessi ja arviointi toteutetaan useita kertoja. Kuva 13 havainnollistaa estimaatin harhaisuutta ja varianssia. [3, luku 7], [41]

## 4.1 Mallin sisäinen validointi

Opetusjoukoksi (*training set*) kutsutaan joukkoa, jota käytetään mallin rakentamiseen. Mallin opettamisen jälkeen on olennaista tietää, kuinka hyvin se toimii ei-opetusjoukon datalla. Sitä varten tarvitaan oma näytejoukkonsa, validointijoukko (*validation set*). Validointijoukon avulla arvioidaan mallin ennustekykyä. [3]




Kuva 14: Mallin arvioinnin prosessi, mukaillen [41]

Mallin arvioinnin tekeminen on esitelty kuvassa 14. Ensin käytössä oleva data jaetaan opetus- ja validointijoukkoihin (vaihe 1). Menetelmiä datan jakamiseen opetus- ja validointijoukkoihin esitellään luvussa 4.2. Datan jakamisen jälkeen valittu oppiva algoritmi opetetaan opetusjoukon näytteillä (vaihe 2), jotta saadaan malli $\hat{f}(X)$. Vaiheessa 3 mallille syötetään validointijoukon syötemuuttujat $X_V$ ja saadaan vasteelle ennuste $\hat{Y_V}$. Vaiheessa 4 ennusteiden $\hat{Y_V}$ ja havaintojen $Y_V$ avulla lasketaan estimaatti ennustekyvylle valituilla ennustekyvyn mittareilla. Estimaatti kannattaa yleensä laskea yhteenvetona lukuisista mittauksista, jotka on tehty hieman eri opetus- ja validointijoukoille. Näin voidaan vähentää satunnaisen vaihtelun merkitystä eli estimaatin varianssia. Tämä tarkoittaa sitä, että mallin arvioinnin prosessin vaiheet

1-4 toistetaan useita kertoja ja näin saaduista ennustekyvyn estimaateista lasketaan lopulta keskiarvo. [41, 42]

## 4.2 Menetelmiä datan jakamiseen

Kuten luvussa 4.1 todettiin, kannattaa ennustekyvyn estimaatti yleensä laskea useita kertoja eri opetus- ja validointijoukoilla varianssin vähentämiseksi. Malli yleensä sopeutuu opetusdataan, jolloin siitä laskettuna ennustekyky on ylioptimistista. Toisaalta, jos validointijoukosta laskettuja ennustekyvyn arvoja hyödynnetään mallin valinnassa, on valitun mallin ennustekyvyn arvio myös ylioptimistinen. Oikeamman ennusteen saa arvioimalla mallia vielä yhdellä riippumattomalla joukolla, eli testijoukolla (*test set*). Testijoukkoon voidaan ottaa esimerkiksi 25% alkuperäisesti aineistosta. Sitä hyödynnetään erityisesti silloin, kun rakennetaan useita malleja, joista pyritään valitsemaan paras. Testijoukkoa hyödyntävää analyysiä ei aina tehdä, joten esiteltävissä menetelmissä käsitellään erilaisia keinoja jakaa data opetus- ja validointijoukkoihin. [3, luku 7]

Yleisesti käytössä olevat menetelmät datan jakamiseen opetus- ja validointijoukkoihin ovat satunnainen osaotanta, ristiinvalidointi ja bootstrap. Näistä menetelmistä on lisäksi useita variaatioita sen suhteen, kuinka suuriin osiin data jaetaan. Eri menetelmillä lasketuilla estimaateilla on erilainen varianssi, harhaisuus ja vaatimukset datalle. Yhteenveto yleisimmistä menetelmistä löytyy taulukosta 8. [42], [3, luku 7]

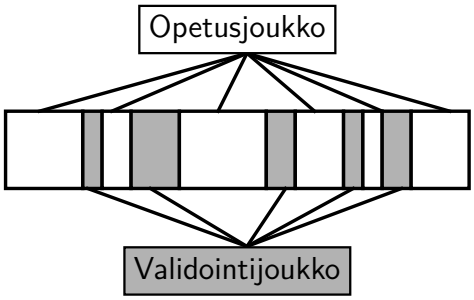
Taulukko 8: Menetelmiä datan jakamiseen. Harhaisuuden kohdalla - tarkoittaa pessimistisyyttä ja + optimistisuutta. Symbolien lukumäärä kuvaa voimakkuutta [42, 3].

| | Harhaisuus | Varianssi |
|---|---|---|
| Satunnainen jako | − − − | +++ |
| Satunnaistettu osaotanta | − − − | ++ |
| 5-ositettu ristiinvalidointi | −− | + |
| 10-ositettu ristiinvalidointi | − | + |
| N-ositettu ristiinvalidointi | 0 | ++ |
| Bootstrap | + + + | + |
| .632 bootsrap | -/++ | + |
| +.632 bootstrap | - | + |

### 4.2.1 Satunnaistettu osaotanta

Jos dataa on käytössä paljon, helpoin keino on jakaa se sattumanvaraisesti kahteen osaan. Usein käytetty jako on $\frac{2}{3}$ opetusjoukkoon ja $\frac{1}{3}$ validointijoukkoon. Hyvä jakosuhde riippuu kuitenkin otoksen ominaisuuksista kuten sen varianssista. Menetelmällä saadaan vain yksi estimaatti, jonka arvo riippuu paljon siitä, millä tavoin

jako opetus- ja validointijoukkoihin on sattunut menemään. Parempi estimaatti saadaan ottamalla useita samankaltaisia satunnaisotoksia ja laskemalla estimaatti niistä saatujen tulosten keskiarvona, mitä havainnollistetaan kuvassa 15 Tätä kutsutaan satunnaistetuksi osaotannaksi (*random subsampling*). Menetelmä ei ole ongelmaton, sillä opetus- ja testijoukko eivät ole riippumattomia vaan toisessa joukossa aliedustettuna oleva luokka on yliedustettuna toisessa. Lisäksi satunnaistettu osaotanta hyödyntää dataa epätehokkaasti, sillä joukon opetukseen hyödynnetään vain osaa datasta. [42]



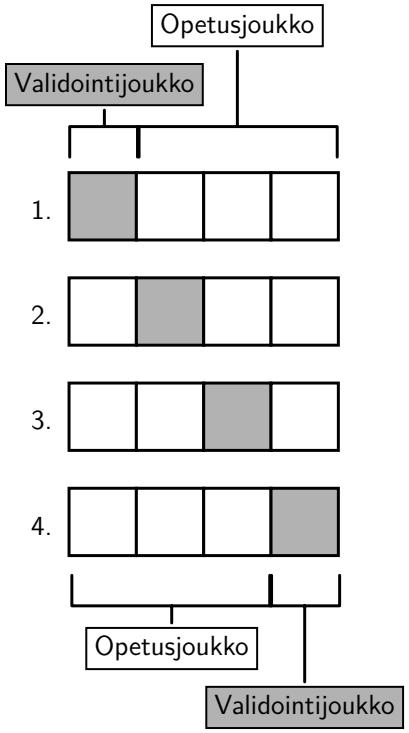

Kuva 15: Satunnaistetussa osaotannassa aineisto jaetaan satunnaisesti opetusjoukkoon (valkoinen) ja validointijoukkoon (harmaa). Jako toistetaan useita kertoja. Mukaillen [42].

Olettaen, että mallin ennustekyky kasvaa opetusjoukon koon kasvaessa, on tällä menetelmällä saatu estimaatti pessimistinen, sillä vain osa datasta on käytetty opettamiseen. Mitä suurempi osa datasta otetaan validointijoukkoon, sitä pessimistisempiä ovat estimaatit. Toisaalta, mitä vähemmän testijoukossa on dataa, sitä suurempi on siitä lasketun mittarin varianssi. Steyerberg ym. [29] tutkivat näiden menetelmien käyttöä mallin sisäiseen validointiin logistisen regression tapauksessa. Heidän tuloksiensa perusteella satunnaista osaotantaa ei tulisi käyttää, ellei otoskoko ole suuri (kummallekin vasteluokalle yli 40 näytettä), sillä se johtaa ennusteen pessimistisyyteen ja estimaatin suureen varianssiin. [42]

### 4.2.2 Ristiinvalidointi

Yksi yksinkertaisimmista ja laajimmin käytössä olevista menetelmistä mittareiden arvojen estimointiin on ristiinvalidointi (*cross-validation*). K-ositetussa *(K-fold)* ristiinvalidoinnissa aineisto jaetaan K yhtä suureen osaan. Osista K-1 kappaletta käytetään mallin opettamiseen, eli ne muodostavat opetusjoukon. Jäljelle jäänyt k:s otosjoukon osa hyödynnetään mallin arviointiin, eli se toimii validointijoukkona. Näin saadulle mallille lasketaan ennustekyvyn estimaatti. Sama toistetaan kaikille osioille $k = 1, 2, ..., K$ ja lopullinen estimaatti saadaan laskemalla näistä keskiarvo. Ristiinvalidointia on havainnollistettu kuvassa 16 arvolla K=4. [3, luku 7]

Jos K=N, kutsutaan menetelmää myös leave-one-out -ristiinvalidoinniksi. Tällöin näyteellä i arvioidaan mallia, joka on opetettu kaikilla muilla näytteillä. Leave-one-out -ristiinvalidoinnissa harhaisuus on matala, mutta varianssi voi olla suuri, sillä kaikki N opetusjoukkoa ovat hyvin samanlaisia. Usein käytetty K:n arvo on 2, 5



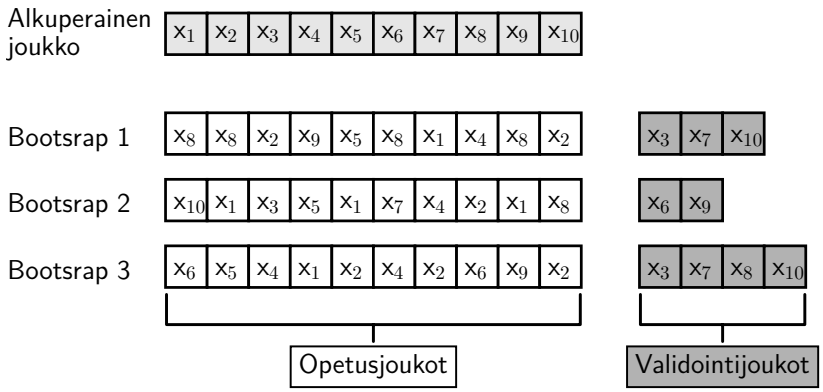

Kuva 16: Esimerkki 4-ositetusta ristiinvalidoinnista. Ristiinvalidoinnissa data jaetaan K osaan, joista yhtä käytetään kerrallaan validointijoukkona ja loppuja K-1 osaa opetusjoukkona. Mukaillen [41].

tai 10. Näille arvoille estimaatin varianssi on pienempi, mutta toisaalta se voi olla pessimistisesti harhainen, sillä opetusjoukon koko on pienempi. Kuinka harhaisen estimaatin ristiinvalidointi antaa, riippuu opettamiseen käytettävän otokseen koon vaikutuksesta ennustetarkkuuteen, mistä kerrotaan enemmän luvussa 4.4. Jos ennustekyky ei laske vaikka otoskokoa pienennetään, saattaa estimaatin harhaisuus olla vähäinen. [3, luku 7], [42]

Ristiinvalidoinnista on olemassa myös tasapainotettu (*stratified*) versio, jossa jokaiseen osioon valitaan näytteet niin, että vastemuuttujia on samassa suhteessa. Esimerkiksi ennustettaessa binääristä vastemuuttujaa otoksessa, jossa keskimäärin 60% vastemuuttujista saa arvon 1, osioidussa ristiinvalidoinnissa jokaiseen osioon pyritään valitsemaan näytteet niin, että noin 60% vastemuuttujista arvo on 1. Monissa tilanteissa tämä menetelmä vähentää estimaatin harhaisuutta ja varianssia. [42]

### 4.2.3 Bootstrap-menetelmä

Bootstrap-menetelmä on yleismenetelmä tilastollisen tarkkuuden arviointiin. Sen perusideana on ottaa aineistosta samankokoisia bootstrap-joukkoja $X_k^*, k = 1, ..., B$ käyttäen yksinkertaista satunnaisotantaa takaisin palauttaen (*simple random sampling with replacement*). Joukkoja otetaan B kertaa ja sama havaintoarvo voi esiintyä bootstrap-joukossa useita kertoja. Laskemalla näille B joukolle jonkin tilastollisen tunnusluvun, voidaan arvioida tunnusluvun otantajakaumaa. [3, luku 7]




Kuva 17: Bootstrap-menetelmässä datasta otetaan saman kokoinen opetusjoukko käyttäen yksinkertaista satunnaisotantaa takaisin palauttaen. Validointijoukkona käytetään niitä alkuperäisen datan näytteitä, jotka eivät kuulu opetusjoukkoon. Jakaminen toistetaan useita kertoja. Mukaillen [41].

Ennustekyvyn arviointiin tätä menetelmää voitaisiin käyttää opettamalla malli bootstrap-joukolle ja validoimalla se alkuperäisellä opetusjoukolla. Näin ei kuitenkaan välttämättä saada hyvää arvioa ennustekyvylle, sillä bootstrap-joukossa ja opetusjoukossa on samoja alkioita, mikä saa aikaan ylioptimistisia arvioita. Menetelmästä on kehitetty versio .632 bootstrap, jossa pidetään kirjaa ennustekyvystä vain niille näytteille, jotka eivät kuulu kyseiseen bootstrap-joukkoon. Tällöin siis opetusjoukko on bootstrap-joukko ja validointijoukko ovat ne näytteet, jotka eivät siihen kuulu. Tätä datan jakamista havainnollistetaan kuvassa 17. Mallin ennustekykyä voidaan tällöin arvioida opettamalla malli opetusjoukolla ja laskemalla ennustekyky validointijoukolla. Sama toistetaan useita kertoja nostamalla uusia bootstrap-joukkoja, ja saaduista arvoista lasketaan keskiarvo, jotta saadaan estimaatti mallin ennustekyvylle. Näin saatu estimaatti on pessimistisesti harhainen ja sen varianssi on kohtalainen. Menetelmä .632 bootsrap on kehitetty pienentämään estimaatin harhaisuutta. Sen antamat arvot saattavat kuitenkin olla joissain tilanteissa hyvinkin optimistisia, etenkin jos malli on ylisovittunut. Menetelmästä on kehitetty myös versio +.632 bootstrap, joka pyrkii ottamaan ylisovittamisen huomioon. [3, luku 7], [42, 29]

## 4.3 Mallin ulkoinen validointi

Ulkoisessa validoinnissa mallia arvioidaan aineistolla, joka on eri populaatiosta kuin mallin rakentamiseen käytetty aineisto. Sen keskeisenä tavoitteena on tutkia mallin ennusteiden yleistyvyyttä [4]. Tyypillisesti mallin ennustekyky on alhaisempi ulkoisen kuin sisäisen validoinnin tapauksessa, mikä voi johtua esimerkiksi erilaisesta populaatiosta tai käytetyistä menetelmistä datan hankinnassa. Yleisiä löytöjä ovat esimerkiksi syötemuuttujista riippumaton ero vastemuuttujassa tai syötemuuttujien erilainen vaikutus. Nämä voivat johtua esimerkiksi erilaisesta muuttujien määrittelystä, mittausmenetelmistä tai tutkittavasta populaatiosta [43]. [29]

Toistettavuutta voidaan tutkia sisäisellä validoinnilla jakamalla dataa opetus- ja validointijoukkoon (luku 4.1), mutta siirrettävyyttä tutkittaessa tarvitaan dataa

toisesta populaatiosta. Ulkoista validointia voidaan jossain määrin tehdä jakamalla data epäsatunnaisesti osiin esimerkiksi sijainnin perusteella [29]. Koskaan ei voi olla täysin varmaa, että malli toimii uudella näytteellä. Mitä erilaisemmilla populaatioilla mallia on tutkittu, sitä varmempana voidaan pitää siirrettävyyttä. Siirrettävyyttä voi tarkastella ainakin ajallisesta, maantieteellisestä, metodologisesta, vastemuuttujan kirjon ja seurantajakson pituuden näkökulmasta. Ajallinen näkökulma liittyy mallin ennustekykyyn eri ajanjaksojen näytteille ja maantieteellinen eri alueiden. Metodologinen siirrettävyys liittyy eroihin muuttujien määrittelyssä ja datan keräämisessä. Vastemuuttujan kirjon tapauksessa kyse on populaatioissa olevista eroista keskimääräisissä arvoissa tai hieman erilaisesta vasteesta, kun taas seurantajakson pituuden tapauksessa kyseessä on lyhyempi tai pidempi seurantajakso. [4]

Justice ym. [4] esittelevät kumulatiivisen menetelmän mallin yleistyvyyden määrittelemiseksi. Sen vaiheet edustavat eri laajuudella tehtyjä validointitutkimuksia. Vaiheet eroavat kahdella tavalla. Mallin arvioimisessa käytetty data voi olla joko samaa tai myöhemmin tai eri alueelta kerättyä kuin opettamisessa käytetty. Riippuen siitä, miten data eroaa, voidaan tutkia ajallista, maantieteellistä tai vastemuuttujan kirjoon liittyvää siirrettävyyttä. Myös tutkimuksen toteuttajat ja aineiston kerääjät voivat olla joko samoja tai eri, jolloin voidaan tutkia metodologista siirrettävyyttä. Menetelmän vaiheet sekä mihin siirrettävyyden näkökulmaan ne voivat vastata, löytyvät taulukosta 9 riveiltä 0-4. Menetelmä on kumulatiivinen, sillä malli voi siirtyä vaiheesta toiseen, kun sitä validoidaan uusissa tutkimuksissa eri tavoin kerätyllä datalla.

Toll ym. [43] jakavat validointitutkimukset ajalliseen (vastaa seurantatutkimusta), maantieteelliseen (vastaa monikeskustutkimusta), ja domain validointiin. Domain validoinnissa tutkitaan ennakointimallin yleistyvyyttä eri kontekstiin, jossa näytteiden populaatiot eroavat. Tämä voi tarkoittaa esimerkiksi vertailua yliopistosairaalan ja terveyskeskuksen tai perusterveydenhuollon ja erikoissairaanhoidon välillä. Domain validointi antaa näistä parhaan kuvan yleistyvyydestä, sillä siinä vaihtelevat eniten metodologiset valinnat ja populaation ominaisuudet. Mitä erilaisemmalla datalla validointia tehdään, sitä vahvempaa on näyttö yleistyvyydestä. Verrattuna Justice ym. [4] menetelmään, domain validaatio on vielä yksi askel kohti monipuolisempaa dataa. Se on viides vaihe taulukossa 9. Yllä esiteltyjen erillisten validointia varten toteutettujen tutkimusten lisäksi siirrettävyyttä voi tutkia hyödyntämällä aiemmin kerättyä dataa. Jos dataa on tarjolla riittävän monipuolisesti, voidaan sen avulla toteuttaa mikä tahansa validaation vaiheista jakamalla näytteet soveltuvalla tavalla.

## 4.4 Otoskoon vaikutus ennustekykyyn

Riittävä otoskoko riippuu esimerkiksi datasta, tutkimuskysymyksestä ja käytetystä mallista.Ennakointimalleissa on yleensä useita muuttujia, mikä vaikeuttaa vaadittavan otoskoon määrittämistä. Jos otoskoko on pieni verrattuna muuttujien määrään, riskinä on mallin alhainen luotettavuus ja ylisovittaminen. Yksi määritelmä on, että lineaarisissa regressiomalleissa tulisi olla vähintään 20 näytettä ja logistisissa regressiomalleissa 10 näytettä jokaista mallissa olevaa syötemuuttujaa kohden [44]. Muissa tutkimuksissa on esitetty myös erilaisia arvioita. Vaikka näytteitä olisi paljon, saattaa

Taulukko 9: Ulkoisen validoinnin tutkimustyyppejä ja niistä saatavat näkökulmat yleistyvyyteen. x=toteutuu aina (x)=toteutuu riippuen tutkimuksesta, tyhjä=ei toteudu.

| | Vaiheen nimi | Joukko | Tutkijat | Toistettavuus | Ajallinen | Maantieteellinen | Metodologinen | Vastemuuttuja |
|---|---|---|---|---|---|---|---|---|
| 0 | Sisäinen | Sama | Sama | x | | | | |
| 1 | Seurantatutkimus | Myöhemmin | Sama | x | (x) | | | |
| 2 | Riippumaton tutkimus | Eri | Eri | x | (x) | (x) | x | (x) |
| 3 | Monikeskus tutkimus | Useilta alueilta | Sama | x | (x) | x | (x) | x |
| 4 | Useita riippumattomia tutkimuksia | Useilta alueilta | Eri | x | (x) | x | x | x |
| 5 | Domain validaatio | Eri domain | Eri | x | x | x | x | x |

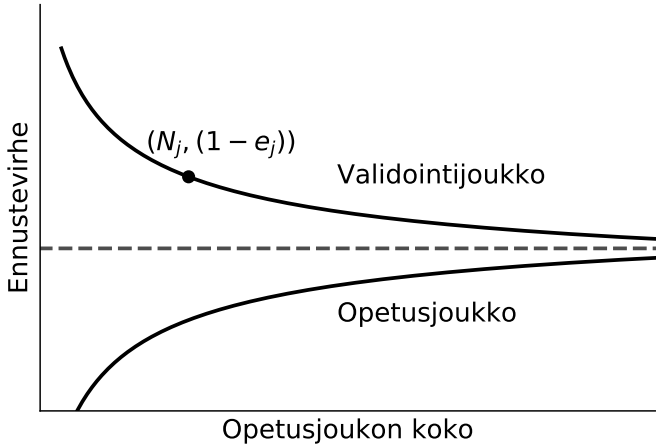
efektiivinen otoskoko (*effective sample size*) olla pieni, jos jokin vastemuuttujan arvo on aliedustettu. [45, 9]

Kaikille malleille ei voida määrittää tällaisia yksinkertaisia rajoja. On kuitenkin huomattu, että monenlaisilla luokittelumalleilla ennustekyky kehittyy potenssilakia noudattaen otoskokoa kasvatettaessa [46]. Oppimiskäyrän avulla voidaan arvioida sitä, miten mallin ennustekyky kehittyisi, jos otosjoukkoa kasvatettaisiin.

### 4.4.1 Oppimiskäyrä

Oppimiskäyrä (*learning curve*) kuvaa opetusjoukon koon $N_O$ vaikutusta mallin ennustekykyyn. Yhteys noudattaa yleensä käänteistä potenssilakia. Oppimiskäyrät voi yleensä jakaa kolmeen osioon. Ensimmäisessä osiossa ennustekyky kasvaa nopeasti otoskoon kasvaessa. Toisessa osiossa kasvu hidastuu merkittävästi. Kolmannessa osiossa malli saavuttaa ennustekyvyn kynnysarvonsa, jonka jälkeen ennustekyky ei kasva, vaikka otoskokoa lisätään. Kuvassa 18 on esimerkki yleisistä oppimiskäyristä, joissa on kuvataan mallin ennustevirhettä opetus- ja validointijoukoille. [47, 48, 46]

Figueroa ym. [46] esittelevät kolmivaiheisen menetelmän luokittelijan ennustekyvyn ennustamiseen. Ensimmäisessä vaiheessa määritetään oppimiskäyrän pisteet, toisessa sovitetaan niihin käyrä ja lopuksi ennustetaan otoskokoa. Oppimiskäyrän pisteet $N_j, e_j$ määritetään opettamalla ja testaamalla luokittelumalli eri kokoisilla opetusjoukoilla. Opetusjoukon kokoa $N_j$ voidaan esimerkiksi porrastaa $k$:n havain-




Kuva 18: Oppimiskäyrä kuvaa mallin ennustekykyä opetusjoukon koon funktiona. Kuvaan piirretty ennustekyvyn sijaan -virhe. Ennustevirhe opetusjoukossa kasvaa ja validointijoukossa pienenee opetusjoukon koon kasvaessa, lähestyen yhteistä rajaarvoa. Mukailtu Cortes ym. [47]

non välein, jolloin $N_j = kj, j = 1, 2, ..m$ ja $km$ on korkeintaan $N_O$. Eri kokoisilla opetusjoukoilla rakennetuille malleille lasketaan ennustekyvyn estimaatti $e_j$ esimerkiksi käyttämällä ristiinvalidointia. Yhtälön 28 mukainen potenssifunktio sovitetaan pisteisiin. Parametrien a, b ja c arvot riippuvat otosjoukosta ja luokittelualgoritmista. Parametri c:n arvot ovat välillä [-1, 0] ja a:n ovat huomattavasti pienempiä kuin yksi. Parametri a edustaa pienintä saavutettavissa olevaa virhettä, b ennustekyvyn kasvun nopeutta ja c ennustekyvyn kasvun hidastumisen nopeutta. Kuten yhtälöstä huomataan, $e_j$ kasvaa asymptoottisesti kohti suurinta saavutettavissa olevaa ennustekykyä, joka näin määriteltynä on (1-a). Sovittamisen voi tehdä esimerkiksi epälineaarisella painotetulla pienimmän neliösumman menetelmällä. Figueroa ym. [46] ehdottavat, että potenssifunktiota sovitettaessa pisteitä $(N_j, e_j)$ painotetaan kertoimilla $\frac{j}{m}$. Oletuksena on, että suurilla opetusjoukoilla opetetun luokittelijan ennustekyky ennustaa tulevaa ennustekykyä paremmin.

$$y_j = (1 - a) - bx^c \tag{28}$$

Viimeisessä vaiheessa sovitettua, yhtälön 28 mukaista käyrää hyödynnetään arvioimalla luokittelun tarkkuutta tai muuta valittua mallin ennustekyvyn mittaria suuremmille otosjoukoille. Lisäksi määritetään 95 % luottamusväli näille arvioille. Tällä menetelmällä voidaan pienellä määrällä dataa arvioida, kuinka suuri opetusjoukon koon tulisi olla, jotta saavutettaisiin riittävä ennustekyky.

# 5 Ennakointimallin tuottaman hyödyn arviointi

Mallien ennustekyvyn ja yleistyvyyden tarkastelu on tärkeää, sillä vääriä tuloksia antava malli on harvoin käyttökelpoinen. Niiden avulla ei kuitenkaan voi arvioida mallin käytön hyötyjä tai kustannuksia, joten mallien valintaan ja arviointiin tarvitaan myös muita työkaluja [6].

Luvussa 5.1 esiteltävä päätösanalyyttinen (*decision analysis*) lähestymistapa on yksi näkökulma mallin hyödyn arviointiin. Siihen liityen esitellään menetelmiä havainnollistaa päätöstilanteita päätöspuun avulla (luku 5.1.1) sekä määrittää hyötyjä ja kustannuksia (luku 5.1.2). Hyötyjen ja kustannusten määrittämistä varten hyödynnetään kustannusvaikuttavuusanalyysiä (*cost-effectiveness analysis)* ja kustannus-hyötyanalyysiä (*cost-benefit analysis*). Kustannusvaikuttavuusanalyysillä voidaan vertailla monipuolisesti erilaisten toimenpiteiden vaikutuksia ja kustannuksia. Siinä vaikutukset lasketaan usein laatupainotettuina elinvuosina ja kustannukset hoitoon käytettyinä rahallisina resursseina. Kustannus-hyötyanalyysi puolestaan pyrkii käsittelemään myös hyötyjä rahallisina. Menetelmät ovat laajoja ja niiden hyödyntäminen vaati laajamittaista hyötyjen ja kustannusten tutkimista ja laskemista, mikä ylittää tämän diplomityön rajauksen. [49, 19]

Mallin hyödyllisyyttä arvioitaessa on hyödyllistä löytää sille vertailukohta, mitä käsitellään luvussa 5.2. Moons ym. [50] ehdottavat mallin käytön vaikutuksen tutkimista vertailemalla tuloksia tilanteissa, joissa malli on ja toisaalta ei ole käytössä. Näin saadaan tietoa mallin vaikutuksista käytössä. Tutkimus voidaan toteuttaa esimerkiksi satunnaistettuina seurantatutkimuksena (*randomized follow-up studies*) tai poikkileikkaustutkimuksena (*cross-sectional studies*). Tämän tyyppinen lähestymistapa vaatii paljon resursseja. Monissa tilanteissa riittääkin verrata mallia jo käytössä oleviin tai triviaaleihin malleihin. Luvussa 5.3 esitellään muutama tapa laskea binäärisen vasteen mallin tuottamia hyötyjä.

Luvussa 5.4 esitellään joitakin kokonaisvaltaisia menetelmiä mallien suorituskyvyn arviointiin. Menetelmät eivät ole kaikenkattavia, mutta niitä hyödynnetään viitekehyksen toteuttamisessa. Esimerkiksi Collins ym. [51] ja Toll ym. [43] suosittelevatkin katsauksissaan lisää tutkimusta arviointimenetelmien määrittelemiseksi.

## 5.1 Päätösanalyyttinen lähestymistapa

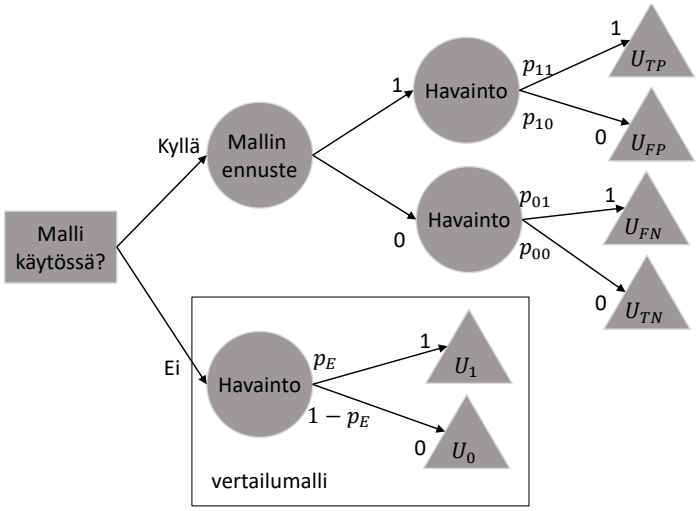
Sosiaali- ja terveysalalla tehtävät päätökset ja päätösketjut ovat usein niin monimutkaisia, että niihin vaikuttavia tekijöitä ei voi kerralla pitää mielessä. Ne myös usein sisältävät vaikeasti hahmotettavaa epävarmuutta ja ristiriitaisia tavoitteita. Näitä haasteita voi ratkoa päätösanalyyttisillä menetelmillä. Päätösanalyysin tärkeänä tehtävänä on auttaa ilmiön ja siihen vaikuttavien muuttujien ymmärtämisessä. Menetelmät mahdollistavat eri vaihtoehtojen tuottaman hyödyn ja kustannusten analysoinnin ja epävarmuuksien tarkastelun esimerkiksi herkkyysanalyysillä [6]. Herkkyysanalyysissä tarkastellaan sitä, miten lähtöoletuksien muuttuminen vaikuttaa mallin tulokseen. [52]

Päätösanalyyttisille menetelmille yhteistä on systemaattinen ja kvantitatiivinen päätösten tarkastelu. Monimutkaiset päätökset jaetaan osiin, osia analysoidaan

erillisinä ja analyysit yhdistetään systemaattisesti avuksi päätöksentekoon. Parhaimmillaan näin voidaan varmistaa, että saadut päätökset seuraavat loogisesti tiedoista ja määritetyistä arvoista. Päätösanalyysin perusperiaate on maksimoida odotettua hyötyä. [52]

Hunink ym. [52] mukaan ensimmäinen vaihe päätösanalyysissä on ongelman ja tavoitteiden määrittely. Pääasiallisen huolen löytämistä voi helpottaa sen pohtiminen, mitä tapahtuisi jos mitään ei tehtäisi. Seurauksia voi kirjata ylös esimerkiksi seuraustauluun (*consequence table*). Ongelmaa kannattaa tarkastella ja uudelleen muotoilla useista näkökulmista, kuten eri ihmisten tai eri tieteenalojen näkökulmasta. Lisäksi keskeiset huolet ja tavoitteet tulee määrittää. Toisessa vaiheessa mietitään mahdolliset päätösvaihtoehdot ja niiden hyödyt ja kustannukset. Vaihtoehtoja voi havainnollistaa esimerkiksi luvussa 5.1.1 esitetyn päätöspuun avulla. Jokaisen päätösvaihtoehdon seuraukset mietitään, ja epävarmojen tapahtumien todennäköisyydet määritetään. Jos erilaisia seurauksia on useampia, vaihtoehdot saattavat olla keskenään ristiriidassa ja eri seurauksien väliset arvotukset täytyy myös määrittää. Viimeiseksi eri vaihtoehtojen hyödyt ja kustannukset yhdistetään ja paras vaihtoehto valitaan. Tätä helpottaa sen pohtiminen, mitkä asiat ovat tärkeimpiä ja missä asioissa vaihtoehdot eroavat merkittävästi. Eri vaihtoehtojen odotettu hyöty lasketaan painottamalla seuraukset niiden todennäköisyyksillä. Lähtökohtaisesti suurimman hyödyn ja/tai pienimmän kustannuksen tuottavat vaihtoehdot valitaan ja epävarmuuksia tutkitaan herkkyysanalyysillä.

### 5.1.1 Päätöspuut




Kuva 19: Päätöspuussa kuvataan päätökset (neliöt), satunnaisuudet (ympyrät) ja niihin liittyvät todennäköisyydet $p_i$ sekä seuraukset (kolmiot) ja niistä syntyvät hyödyt ja kustannukset $U_i$. Päätöspuu kulkee ajallisessa järjestyksessä vasemmalta oikealle.

Päätöspuu on keino havainnollistaa päätöksentekotilanteita. Sitä ei pidä sekoittaa

samannimiseen luokittelualgoritmiin. Kuvassa 19 on esimerkki päätöspuusta, jolla kuvataan binäärisen luokittelumallin käyttöönottoa. Päätöspuissa piirretään graafi, joka kulkee loogisessa ajallisessa järjestyksessä vasemmalta oikealle. Se koostuu päätössolmuista (*decision node*), sattumasolmuista (*chance node*) ja seuraussolmuista (*consequence node*) sekä niitä yhdistävistä kaarista (*arcs*).

Ennakointimallien tapauksessa tutkittava päätös on se, kannattaako mallia käyttää. Oletetaan havainnollistuksen vuoksi malliksi binäärinen luokittelu, jossa pyritään ennustamaan sairastuuko potilas. Puussa päätöksenä on se, otetaanko malli käyttöön vai ei, jolloin vertaillaan tutkittavaa mallia ja triviaalia tai olemassa olevaa mallia. Päätössolmua (*decision node*) merkitään kuvassa neliöllä. Kuvassa 19 alemmassa haarassa mallia ei oteta käyttöön, jolloin lopputilanne määräytyy vertailumallin perusteella. Kuvan tilanteessa vertailumalli on potilaiden sattumanvarainen luokittelu esiintyvyyden $p_E$ perusteella. Päätöspuissa sattumasolmuja kuvataan ympyröillä ja niissä kaari valitaan sattumanvaraisesti tiettyjen, joko tunnettujen tai tuntemattomien, todennäköisyyksien pohjalta. Kuvassa oikealla on kolmioilla kuvattavat, puun 'latvassa' sijaitsevat seuraussolmut (*consequence node*), joihin päädytään, kun koko puu on kuljettu läpi. Seuraussolmuille on määritettävissä hyödyt ja kustannukset $U_i$. Hyötyjen ja kustannusten määrittelystä on lisää luvussa 5.1.2. [52, 53]

Päätöspuun haarassa, jossa malli otetaan käyttöön, on lisäksi sattumasolmu mallin ennuste. Esimerkkimallin tapauksessa havainto 1 tarkoittaa, että potilas sairastuu ja nolla, että potilas ei sairastu. Riippuen ennusteesta, näytteiden todennäköisyys saada arvo 1 eroaa. Esimerkiksi $p_{11}$ tarkoittaa todennäköisyyttä havaintoon 1, jos mallin ennuste on 1. Päätöspuun avulla voi valita parhaat vaihtoehdot kulkemalla ensin puuta seuraussolmuista lähtien oikealta vasemmalle. Sattumasolmuissa lasketaan hyötyjen ja kustannusten odotusarvo. Päätössolmuissa taas valitaan se kaari, jonka odotusarvo on suurin. Kun koko puu on näin kuljettu läpi, tiedetään mikä päätösten ketju tuo parhaat hyödyt ja kustannukset. Näin voidaan verrata hyötyjä, jotka on saavutettavissa mallilla ja ilman sitä. [53]

### 5.1.2 Tavoitteet, hyödyt ja kustannukset sosiaali- ja terveysalalla

Hunink ym. [52] mukaan terveydenhuollon päätavoite on vähentää sairauksien vaikutuksia, eli joko parantaa terveyttä tai hidastaa terveyden laskua. Se voi tarkoittaa esimerkiksi sairauden estämistä, parantamista, hidastamista tai oireiden lievittämistä. Tavoitteita määriteltäessä tulisi erotella välinetavoitteet (*means objectives*) ja perustavaa laatua olevat tavoitteet (*fundamental objectives*). Välinetavoitteisiin pyritään, jotta saavutettaisiin perustavaa laatua olevia tavoitteita. Eroa näiden kahden väillä voi löytää kysymällä miksi, jolloin liikutaan aina lähemmäs perustavaa laatua olevia tavoitteita. Esimerkki perustavaa laatua olevasta tavoitteesta on potilaan elämän laadun parantaminen, jota voidaan saavuttaa pienentämällä riskiä sairastua tai hoitamalla oireita. Taustalla olevien tavoitteiden ymmärtäminen auttaa löytämään useampia välinetavoitteita. [49, 52]

Terveydenhuoltoon liittyvissä sovelluksissa hyödyt ja kustannukset liittyvät usein ihmisten terveydentilaan. Ne voivat koskea esimerkiksi sairastavuutta, kuolleisuutta, elämänlaatua tai oireiden vakavuutta. Toisaalta ne voivat liittyen resurssien

tehokkaaseen käyttöön, kuten lääkkeiden, toimenpiteiden, laitteiden käytön tai hoitohenkilökunnan työaikoihin. Myös esimerkiksi kuljetuksesta ja tilojen käytöstä aiheutuu kustannuksia. Lisäksi potilaan ajankäyttö on mitattava asia. [49]

Hyödyt ja kustannukset voivat olla välillisiä tai välittömiä. Välittömillä kustannuksilla ja hyödyilllä tarkoitetaan palveluiden, tuotteiden tai muiden toimenpiteeseen kulutettavien resurssien arvoa. Välillisten kustannusten ja hyötyjen arvo puolestaan on epäsuoraa, se voi esimerkiksi tarkoittaa sairaalassa olevan potilaan lasten hoitokustannuksia tai sairastuneen työn tuottavuuden laskua. [49]

## 5.2 Vertailumallin määrittäminen

Mallia voi verrata triviaaliin, olemassa olevaan ($f_0$) tai parhaaseen mahdolliseen malliin $f_{perfect}$. Triviaali malli voi esimerkiksi tarkoittaa sitä, että potilaat luokitellaan sattumanvaraisesti korkean riskin todennäköisyydellä $p_E$, joka on taudin esiintyvyys. Toisaalta, se voi tarkoittaa mallia, jossa epäillään sairautta kaikilla ($f_{all}$), tai ei kenelläkään $f_{none}$ [54], [6]. Paras mahdollinen taas tarkoittaa hypotettista mallia, joka antaa aina oikeita tuloksia. Esimerkiksi binäärisen vasteen tapauksessa täydellinen malli on täysin kalibroitu, sen oikeiden positiivisten osuus on 1 ja väärien positiivisten osuus on nolla 0, eli mallin sensitiivisyys ja spesifisyys ovat 1 [6]. [19]

Jos tilanteeseen on olemassa jo jokin käytössä oleva ja eksplisiittinen ennakointimalli, on vertailukohdan löytäminen helpompaa. Usein näin ei kuitenkaan ole. Vaikka mallia ei olisi, ei ennakointi välttämättä ole täysin sattumanvaraista, vaan esimerkiksi terveydenhoidon ammattilaisten kokemukseen pohjautuvia päätöksiä. Tällöin todellisen vertailumallin määrittäminen voi olla vaikeaa. [55]

Esimerkiksi tilanteessa, jossa ennakointimallin avulla pyritään löytämään ne kotihoidon potilaat, jotka seuraavan vuoden aikana joutuvat ympärivuorokautiseen hoitoon, jotta heille voitaisiin kohdistaa erityistoimenpiteitä, voidaan määritellä seuraavat vertailumallit. Triviaaleina malleina $f_{all}$ tarkoittaa että kaikille kohdistetaan erityistoimenpiteitä, $f_{none}$ että kenellekään ei kohdisteta erityistoimenpiteitä ja $f_{pE}$ että potilas valitaan sattumanvaraisesti todennäköisyydellä $p_E$. Paras mahdollinen malli $f_{perfect}$ löytäisi kaikki hoitoon joutuvat, mutta ei määrittelisi riskiryhmään ketään, joka ei joudu. Olemassa oleva malli $f_0$ taas voi olla esimerkiksi jo olemassa oleva päätössääntö, numeerinen malli tai kotihoidon henkilökunnan epäformaalisti tai tiedostamatta tekemä arvio.

## 5.3 Laskennallisia keinoja hyödyn määrittämiseen

Binäärisen vasteen malleille on kehitetty joitakin laskennallisia menetelmiä mallin tuottaman hyödyn arviointiin. Odotetun hyödyn (*expected utility*) ja nettohyödyn (*net benefit*) tapauksissa binäärisen vasteen ennusteen perusteella päätetään interventiosta. Eri lopputulemiin (TP, FP, TN, FN) liittyy hyötyjä $U_{xx}$, jotka voivat olla niin positiivisia kuin negatiivisia (kustannuksia). Ennusteen tekoon liittyy kustannus $U_{test}$, joka voi olla esimerkiksi jonkin lääketieteellisen testin kustannus.

### Odotettu hyöty

Baker ym. [6] ovat kehittäneet menetelmän, joka käyttää hyötyjä ja kustannuksia valitakseen parhaan hoitovaihtoehdon binäärisen vasteen testin tapauksessa. Se perustuu tietylle mallin positiiviseksi tai negatiiviseksi määrittävälle raja-arvolle $p_t$, joille odotettu hyöty voidaan laskea kaavalla:

$$\begin{aligned} EU_{pt} = p_E\frac{N_{TP,pt}}{N}U_{TP} + p_E(1-\frac{N_{TP,pt}}{N})U_{FN} + (1-p_E)(1-\frac{N_{FP,pt}}{N})U_{TN} \\ + (1-p_E)(\frac{N_{FP,pt}}{N})U_{FP} + U_{test}. \end{aligned} \tag{29}$$

Tässä $p_E$ on esiintyvyys, $N_{TP,pt}$ on oikeiden positiivisten määrä raja-arvolla $p_t$, $U_{TP}$ oikeista positiivisista syntyvät hyödyt ja kustannukset (muille lopputulemille vastaavasti) ja $U_{test}$ on ennusteen tekemisen kustannus.

$U_{perfect}$ tarkoittaa täydellisen ennusteen hyötyä. Täydellisellä mallilla oikeiden positiivisten osuus on 1, väärien positiivisten osuus on 0 ja $U_{test} = 0$. Täydellisen ennusteen hyödyn voi laskea sijoittamalla nämä yhtälöön 29, jolloin saadaan $U_{perfect} = p_E(U_{TP} - U_{FN}) + p_tU_{FN} + (1-p_t)U_{TN}$. Hyöty sille, että ketään ei hoideta, on $U_{none} = p_tU_{FN} + (1-p_t)U_{TN}$ ja että kaikki hoidetaan, on $U_{all} = p_tU_{TP} + (1-p_t)U_{FP}$ [6]. Näitä hyödyntämällä voidaan määrittää kliinisen informaation odotusarvo (net expected value of clinical information) $U_{pt} - U_{none}$ ja täydellisen kliinisen informaation odotusarvo (expected value of perfect clinical information) $U_{perfect} - U_{none}$.

### Nettohyöty

Hyöty, joka saadaan oikein luokitelluista näytteistä on usein hyvin erilainen kuin kustannukset, joita syntyy vääristä positiivisista. Nettohyöty pyrkii ottamaan tämän huomioon, sillä se painottaa oikeita ja vääriä positiivisia niiden suhteellisilla kustannuksilla. [54]

Nettohyöty kuvaa tilannetta, jossa tehdään päätös interventiosta perustuen malliin tai testitulokseen. Oletetaan, että väärille positiivisille ja negatiivisille on kustannukset $U_{FP}$ ja $U_{FN}$. Lisäksi määritetään raja-arvotodennäköisyys $p_t$, joka kuvaa sitä, millä mallin antamalla todennäköisyydellä päätöksentekijä on epävarma siitä, kannattaako interventio. Tällä todennäköisyydellä väärien positiivisten ja negatiivisten kokonaiskustannukset ovat yhtä suuret. Jos esimerkiksi sairastumistodennäköisyys 10% on se kynnysarvo, jolla väärien positiivisten ja väärien negatiivisten aiheuttamat kustannukset ovat tasapainossa, niin $p_t$ on 10%. [19]

Nettohyöty on kaavalla 30 riskimittareille laskettava mittari, jonka arvo vaihtelee välillä $[-\infty, p_E]$. Hyvällä mallilla on korkea nettohyöty.

$$\mathrm{NB} = \frac{N_{TP}}{N} - \frac{N_{FP}}{N}(\frac{p_t}{1-p_t}) \tag{30}$$

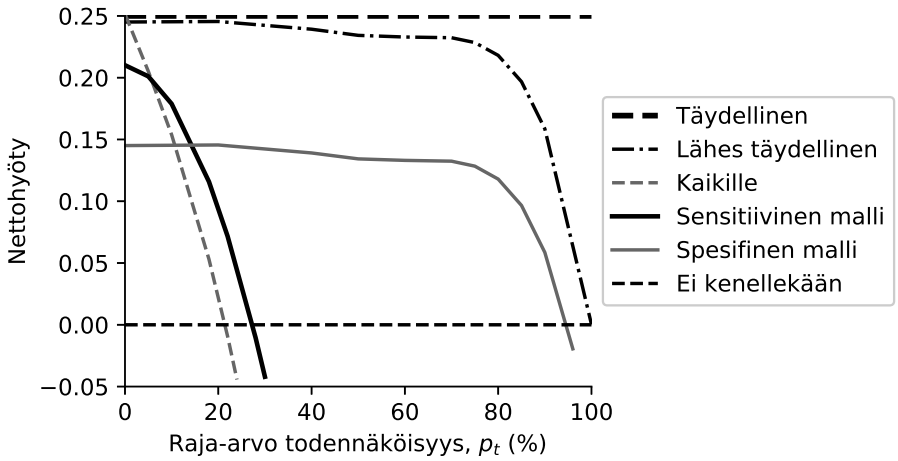
Mallia voidaan verrata kahteen triviaaliin tilanteeseen. Toinen näistä on ettei ketään hoideta, jolloin NB on nolla, koska ei ole oikeita tai vääriä positiivisia. Jos taas päätetään hoitaa kaikki, on $NB_{all} = \frac{p_E - p_t}{1-p_t}$, missä $p_E$ on esiintyvyys. Näin

nettohyödyn voidaan tulkita kuvaavan kasvua oikeiden positiivisten määrässä ilman että väärien positiivisten määrä kasvaa, verrattuna vaihtoehtoon ettei ketään hoideta. Jos kahta mallia vertaillaan, voidaan ero nettohyödyssä laskea kaavalla $\Delta NB = \frac{1}{N}(\Delta N_{TP} - \frac{p_t}{1-p_t}\Delta N_{FP})$. [54]

### Päätöskäyräanalyysi

Monien päätösanalyyttisten menetelmien huono puoli käytännön kannalta on, että ne vaativat dataa mallin rakentamiseen hyödynnetyn otoksen ulkopuolelta. Tähän kuuluu esimerkiksi tietoa kustannuksista tai eri terveydellisten tilojen arvosta. Monet menetelmät myös vaativat, että ennakointimallin vaste on binäärinen. [19]

Päätöskäyräanalyysi (*decision curve analysis*) pyrkii vastaamaan näihin haasteisiin siinä tilanteessa, kun mallia hyödynnetään riskin arviointiin ja korkeariskisille potilaille suoritetaan interventio. Analyysissä tärkeässä osassa on nettohyöty 5.3. Nettohyödyssä olevan raja-arvon $p_t$ määrittäminen on haastavaa, joten päätöskäyräanalyysissä nettohyöty lasketaan useille $p_t$ arvoille ja arvoista piirretään kuvaaja. Samaan kuvaajaan lisätään käyrät myös mahdolliselle verrokkimallille sekä triviaaleissa tapauksissa, joissa joko kaikille toteutetaan ($f_{all}$) tai kenellekään ei toteuteta ($f_{none}$) interventio. Kuvaajaa tarkastellaan relevanteilla $p_t$ arvoilla, ja jos mallin käyrä on ylempänä kuin vertailukohta, siitä voidaan tulkita olevan hyötyä. Esimerkki teoreettisista päätöskäyristä on kuvassa 20. Kuvaan on hahmoteltu käyrät täydelliselle mallille, lähes täydelliselle, sellaiselle jossa kaikille suoritetaan interventio, mallille joka on sensitiivinen (99% sensitiivisyys ja 50% spesifisyys), mallille joka on spesifinen (99% spesifisyys ja 50% sensitiivisyys) ja tilanteessa jossa kenellekään ei suoriteta interventiota. [19]




Kuva 20: Esimerkki päätöskäyrästä erilaisille teoreettisille malleille tilanteessa, jossa esiintyvyys on 25%. Mukailtu [19]

**Suhteellisen hyödyn käyrät**

Baker ym. [6] ehdottavat päätöskäyräanalyysiin laajennusta, jota he kutsuvat suhteellisen hyödyn käyriksi (*relative utility curves*) (yhtälö 31). Ne kertovat, kuinka paljon ennakointimalli kontribuoi hyötyyn verrattuna täydelliseen ennusteeseen. Suhteellisen hyödyn käyrä lasketaan vertaamalla mallin ennusteen antamaa maksimihyötyä (verrattuna mallittomaan tilanteeseen) täydellisen ennusteen hyötyyn (verrattuna mallittomaan tilanteeseen).

$$RU(p_t) = \begin{cases} \dfrac{U_{pt} - U_{All}}{U_{perfect} - U_{All}}, & p_t < p_E. \\ \dfrac{U_{pt} - U_{None}}{U_{perfect} - U_{None}}, & p_t \geq p_E. \end{cases} \tag{31}$$

Tässä malliton tilanne tarkoittaa joko mallin $f_{All}$ tai $f_{None}$ käyttämistä. Jos mallin puuttuessa on hyödyllistä antaa hoitoa kaikille, eli $U_{All} > U_{None}$, on relevantti alue $p_t < p_E$. Vastaavasti relevantti alue on $p_t \geq p_E$, jos $U_{All} \leq U_{None}$.

## 5.4 Esimerkkejä tavoista arvioida malleja

Toll ym. [43] katsauksen mukaan kirjallisuutta on paljon sosiaali- ja terveysalan ennakointimallien kehittämisestä ja tuloksista, mutta mallien toiminnan ja käytön vaikutuksien arvoinnista huomattavasti vähemmän. Collins ym. [51] tutkimuksen mukaan vain murto-osa ennakointimalleja käsittelevistä artikkeleista arvioi mallien suorituskykyä. Tähän lukuun on koottu löydettyjä menetelmiä mallien arviointiin.

Kuten luvussa 2 todettiin, useimmat mallit opetetaan optimoimalla jotain tiettyä mittaria, jolloin ennustekyky kyseisellä mittarilla saattaa olla korkeampi kuin muilla. Toisaalta, jos mallia parannetaan joidenkin mittareiden suhteen, saattavat sen tulokset huonontua toisilla mittareilla mitattuna [56]. Kokonaiskuvan saamiseksi mallin arviointia tulisikin tehdä useilla mittarilla [20]. Esimerkiksi luokittelijat, jotka maksimoivat osajoukon tarkkuutta, toimivat huonosti Hammingin etäisyyden näkökulmasta, ja toisin päin [12].

Monet ennakointimallien arviointiin kehitetyt tavat yhdistelevät eri näkökulmasta mallia tarkastelevia mittareita ja analyysimenetelmiä. Steyerberg ym. [5] mukaan erottelun ja kalibraation tutkiminen on aina tärkeää ennakointimalleille. Lisäksi, jos malli vaikuttaa päätöksentekoon, tulisi tehdä päätösanalyysiä. Lisäksi sovelluksesta riippuen myös muunlaiset menetelmät, kuten NRI, voivat olla hyödyllisiä.

Steyerberg ym. [31] menetelmässä mallin validointiin tarkastellaan kalibrointia, erottelua ja kliinistä hyötyä (*clinical usefullness*). Kalibraation tarkasteluun piirretään kalibraatiokuvaaja ja lasketaan suuren mittakaavan kalibrointi ja kalibraation kulmakerroin. Hosmer-Lemeshown mittaa ei suositella. Erottelun tarkastelua varten piirretään validaatiokuvio ja lasketaan yhteensopivuusindeksi. Kliinistä hyötyä he suosittelevat arvioimaan päätöskäyräanalyysillä. Toisaalta tutkimuksessa Steyerberg ym. [29] logistisen regression tapauksessa tärkeiksi mittareiksi valittiin Nagelkerke $R^2$, Brierin pistemäärä, yhteensopivuusindeksi sekä kalibraatiokuvaajan kulmakerroin.

Collins ym. [51] suosittelevat aina tarkastelemaan kalibrointia ja erottelua, ja pitävät hyödyn tarkastelua hyvin tärkeänä. Heidän tarkastelemansa menetelmät ovat kalibraatiokuvio, Hosmer-Lemeshow test, yhteensopivuusindeksi, ROC-käyrä, Brierin pistemäärä, $R^2$, päätöskäyräanalyysi ja suhteellisen hyödyn käyrät. He suosittelevat ennemmin kalibraatiokuviota kuin Hosmer-Lemeshow -testiä.

Justice ym. [4] suosittelevat mallin arviointia ennustekyvyn (kalibraatio ja erottelu) ja yleistyvyyden (toistettavuus ja siirrettävyys) näkökulmista. Mallin ennustekyky on edellytys mallin yleistyvyydelle. Koska erottelun ja kalibraation suhteellinen tärkeys riippuu sovelluskohteesta, niitä tulisi aina tutkia. Kalibraatio on tärkeää esimerkiksi silloin, kun tavoitteena on arvioida tietyn sairaalan tuloksia, kun taas erottelu on tärkeämpää, jos tavoitteena on suunnata hoidon resursseja potilaille ennustetun riskin perusteella.

# 6 Viitekehyksen määrittäminen

Diplomityön tavoitteena on määrittää viitekehys, jonka avulla erilaisten ennakointimallien suorituskykyä voidaan arvioida monipuolisesti ja kattavasti ja näin varmistaa mallin olevan riittävän hyvä käyttöönotettavaksi. Alla on lyhyesti kuvattu ennakointimallien kehittämisen prosessi, jota diplomityössä kehitetään arvioinnin näkökulmasta [57].

1. Määritetään vastemuuttuja ja laaja lista mahdollisia syötemuuttujia hyödyntäen asiantuntijoiden haastatteluja ja työpajoja.

2. Lähtödata kerätään useista lähteistä, kuten eri toimijoiden asiakas- ja potilastietojärjestelmistä tai tietoaltaista.

3. Dataa esikäsitellään esimerkiksi käsittelemällä puuttuvat ja virheelliset arvot ja normalisoimalla muuttujia. Muuttujia esivalitaan perustuen muuttujien välisiin riippuvuuksiin ja muuttujakohtaiseen ennustetarkkuuteen.

4. Tehokkaimmat syötemuuttujien joukot valitaan muuttujavalintamenetelmillä. Malli ja lopullinen syötemuuttujien joukko valitaan erillisiin datajoukkoihin perustuvalla analyysillä.

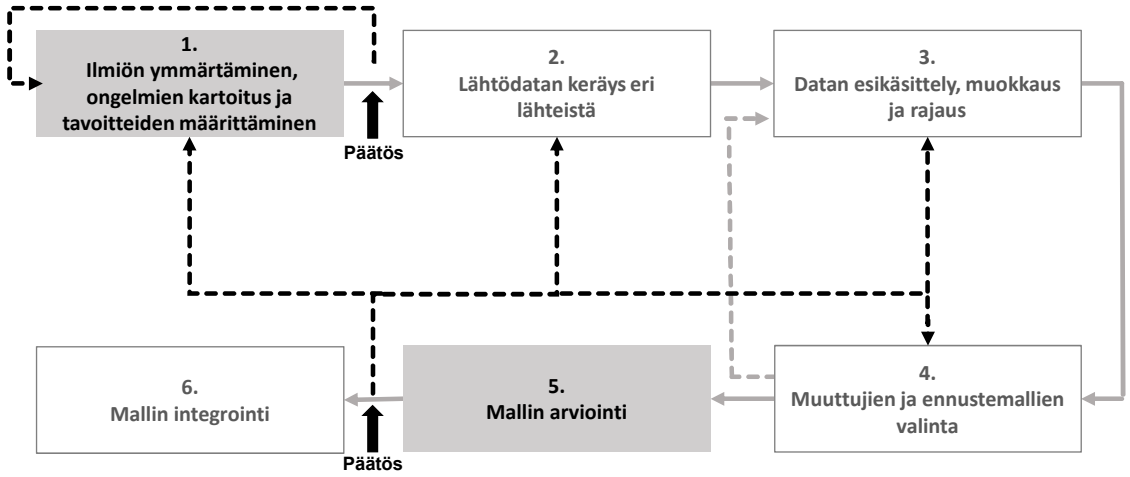
5. Mallin suorituskykyä arvioidaan ulkoisen validaation ja asiantuntijoiden avulla

6. Kehitetty malli toteutetaan tuotantojärjestelmiin ja integroinnin onnistumista arvioidaan piloteilla.

Arviointiin keskittyvä viitekehys koskee erityisesti prosessin vaiheita 1 ja 5. Vaiheet 2-4 ovat luonnollisesti keskeisiä mallin suorituskyvyn kannalta ja niissä jokaisessa tehdään runsaasti valintoja, jotka vaikuttavat niin mallin ennustekykyyn, yleistyvyyteen kuin hyötyyn. Toisaalta vaiheen 6 onnistuminen määrittää mallin arvon käytössä. Näistä jokainen on oma iso kysymyksensä, jotka on kuitenkin rajattu pois tästä diplomityöstä.

Päivitetty ennakointimallien kehittämisen prosessi on esitetty kuvassa 21. Prosessin vaiheet, joita ei tässä diplomityössä muutettu, on esitetty vaaleammalla värillä. Tässä luvussa kuvataan tarkemmin arvioinnin kannalta kehitetyt osat prosessista eli viitekehys sosiaali- ja terveysalan ennakointimallien arviointiin. Yhteenveto viitekehyksestä on liitteessä B. Luvussa 6.1 kuvataan prosessin vaihetta yksi, jota kehitettiin tukemaan laajemmin ilmiön ymmärtämistä, ongelmien määrittämistä ja valintaa, tavoitteiden asetantaa sekä päätöksentekoa mallin toteuttamisen kannattavuudesta. Prosessin vaiheen viisi päivitys on kuvattu luvussa 6.2. Sitä laajennettiin koskemaan ennustekyvyn ja yleistyvyyden tavoitteiden saavuttamista, otoskoon riittävyyden tarkastelua ja mallin käytön hyötyjen ja riskien määrittämistä. Lisäksi tärkeänä osana on kokonaisarviointi, jossa kaikki yllä mainitut alueet otetaan huomioon. Kokonaisarvion pohjalta tehdään päätös mallin tulevaisuudesta.

Lopulliseen viitekehykseen päädyttiin kirjallisuuskatsauksen (luvut 2-5) ja teemahaastattelujen [7, 55, 58, 59] perusteella. Haastatteluissa pureuduttiin viitekehyksen




Kuva 21: Päivitetty ennakointimallien kehittämisen prosessi. Diplomityössä kehitetyt osiot on esitetty tummemmilla teksteillä ja nuolilla. Vaaleammilla teksteillä ja nuolilla kuvattuja osioita ei ole diplomityössä päivitetty. [57]

kannalta tärkeisiin aiheisiin, joista ei saatu riittävästi tietoa kirjallisuuskatsauksesta. Avoimeksi jääneisiin kysymyksiin etsittiin haastateltavat, joilla oli kyseisestä aiheesta kokemusta ja vaadittavaa asiantuntemusta. Haastattelutyypiksi valittiin teemahaastattelu, jotta haastatteluissa voitiin keskittyä viitekehyksen kannalta olennaisiin aiheisiin, mutta antaa haastateltaville tilaa jakaa osaamistaan. Haastatteluissa apuna käytetyt kysymykset ovat liitteessä A.

## 6.1 Vaihe 1: Ilmiön ymmärtäminen, ongelmien kartoitus ja tavoitteiden määrittäminen

Onnistuneen ja vaikuttavan mallin kannalta keskeistä on ilmiön syvällinen ymmärtäminen, keskeisten ongelmien löytäminen ja realististen tavoitteiden määrittäminen jo mallin kehitystyön alussa [55]. Siinä tarvitaan sekä sovellusalueen että mallintamisen asiantuntemusta. Vaihe ei ole lineaarinen vaan ennemminkin iteratiivinen. Esimerkiksi kun mallille asetetaan tavoitteita, saatetaan huomata, ettei ilmiötä ole ymmärretty riittävän syvällisesti. Vaiheen lopputulemana on muodostettu syvällinen käsitys ilmiöstä ja havainnollistettu sitä, valittu ongelmien ja niiden merkittävyyden kartoittamisen perusteella relevantti kehityskohde ja suunniteltu mallin toteutus karkealla tasolla. Lisäksi on valittu tärkeimmät ennustekyvyn ja hyödyn mittarit ja määritelty niille hyväksyttävät arvot sekä arvioitu onnistuneen mallin tuottamaa hyötyä. Kun riittävän syvällinen kuva on muodostettu, tehdään päätös mallin kehitystyön jatkamisesta. Kuvatut asia on selitetty auki tässä luvussa ja tiivistetty liitteeseen B.

### 1.A: Ilmiön ymmärtäminen

Ilmiön ymmärtäminen vaatii keskusteluja ja muuta tiedonvaihtoa sovellusalueen asiantuntijoiden kanssa. Olennaiset selvitettävät asiat riippuvat sovelluskohteesta

ja siitä, kuinka hyvin ilmiö tunnetaan etukäteen. Vaiheen aikana tulee hankkia ymmärrys ilmiön keskeisimmistä ominaisuuksista, joihin kuuluvat esimerkiksi:

- Keskeiset toimijat ja toimenpiteet
- Potilasvirrat
- Informaatiovirrat
- Kohdat, joissa ilmiöön on mahdollista vaikuttaa

Ilmiöstä tulee myös ymmärtää eri osien väliset *vuorovaikutukset* ja keskeiset *dynamiikat*, jotta voidaan ymmärtää osien muodostaman kokonaisuuden toimintaa. Kokonaisvaikuttavuuden vuoksi tulee varmistaa, että koko *hoitoketju* on riittävällä tasolla ymmärretty ja kuvattu, jotta keskitytään kokonaisuuden kannalta merkitseviin kohtiin eikä osaoptimoida. Ilmiötä kannattaa havainnollistaa erilaisilla prosessikaavioilla [60].

Lisäksi ennakointimallin kannalta olennaista on selvittää ilmiöstä olevan datan saatavuus. Tähän soveltuvia kysymyksiä ovat esimerkiksi, mitä asioita ilmiöön liittyen kirjataan ja missä vaiheessa sekä kuka kirjaa ja mihin järjestelmään. Relevanttia tietoa kirjataan usein laajemminkin kuin vain kyseiseen ilmiöön liittyen. Esimerkiksi potilaista saatetaan tietää heidän fyysisiä ominaisuuksiaan tai käyntihistoriaansa. Kannattaakin selvittää myös, millaista muuta kuin ilmiöön liittyvää tietoa on saatavilla. Lisäksi on olennaista tietää, kuinka usein data päivittyy saataville ja kuinka paljon ja kuinka pitkältä ajalta on historiadataa. [55, 61]

### 1.B: Ongelmien kartoitus ja valinta

Ongelmien kartoitukseen tarvitaan sovellusalueen asiantuntemusta. Jotta ongelmia löydetään monipuolisesti, kannattaa keskusteluissa hyödyntää kolmijakoa potilas-yksikkö-hoitoketju. Potilaan näkökulma tuo keskusteluihin mukaan yksittäisen asiakkaan kannalta olennaiset ongelmat ja yksikkönäkökulma palvelutuottajan sekä yksikön johdon näkökulman. Sosiaali- ja terveysalalla yksikkö saattaa usein hoitaa vain osan potilaan kokonaistarpeesta, jolloin hoitoketjunäkökulmaa tarvitaan kokonaisuuden tarkasteluun ja osaoptimointiriskin pienentämiseen. Löydettyjen ongelmien merkittävyyttä vertaillaan myöskin kolmijaon avulla. Tässä vaiheessa olennaisinta on löytää ongelmien kokoluokka, joten merkittävyyttä voi arvioida esimerkiksi asteikolla 1-3, joissa luokka 1 tarkoittaa kymmmeniä tuhansia, 2 satoja tuhansia ja 3 miljoonia euroja. Ongelmien ja niiden merkittävyyden kartoittamista voidaan tehdä esimerkiksi työpajoissa. [55]

Lopulta tehdään valinta siitä, minkä ongelman ratkaisuun keskitytään. Valintaa tehdessä huomioitavia näkökulmia ovat ainakin [55]:

- Ongelma on niin merkittävä, että mallilla voidaan saada riittäviä hyötyjä aikaiseksi
- Ennakointimalli on järkevä ja tehokas tapa ratkaista ongelmaa, eli yksinkertaisemmilla keinoilla tuskin päästäisiin yhtä hyvään lopputulokseen

- Ilmiöön liittyvien toimijoiden kehityspanoksia on laajemmin suunnattu ongelman ratkaisuun, jolloin on mahdollisuus suurempiin ja vaikuttavampiin muutoksiin
- Hyvää dataa on saatavilla riittävästi ongelmaan liittyvän mallin toteuttamiseen
- Onnistuneella ennakointimallilla on aito mahdollisuus vaikuttaa ongelmaan, eli ennusteilla voidaan muuttaa toimintatapoja paremmiksi

### 1.C: Mallin toteutuksen suunnittelu

Ongelman valinnan jälkeen tulee määrittää karkealla tasolla mallin toteutus. Siihen kuuluvat esimerkiksi seuraaviin kysymyksiin vastaaminen [55]:

- Mitä halutaan ennustaa, eli mikä on mallin vastemuuttuja? Eri sovelluskohteissa keskeisin ennustettava asia voi olla hyvinkin erilainen, kuten potilaan sairastumistodennäköisyys tai osaston päivittäinen resurssitarve.
- Mitkä syötemuuttujat ovat potentiaalisimpia vastemuuttujan ennakoinnissa?
- Kuka mallia tulisi käyttämään ja mihin järjestelmään se integroitaisiin?
- Millainen muutos toimintaan mallin avulla halutaan saavuttaa?

### 1.D: Tavoitteiden määrittely

Tavoitteiden määrittely kannattaa aloittaa pohtimalla, mikä mallin ennustekyvyssä on olennaisinta. Halutaanko esimerkiksi välttää vääriä negatiivisia vai saavuttaa keskimäärin oikeita tuloksia? Selvittämisessä voi auttaa sen pohtiminen, millaisista mallin virheistä syntyy eniten kustannuksia tai miten täydellinen malli toimisi. Ennakoinnin kontekstissa myös määritetään, mikä on olemassa oleva tai muu vertailumalli (luku 5.2), jonka suhteen rakennettavaa mallia arvioidaan. Lisäksi päätetään, kuinka paljon vertailumallia paremmin kehitettävän mallin tulisi toimia.

Luvussa 3 esitellyistä ennustekyvyn mittareista valitaan sovelluskohteelle parhaat mallintamisen asiantuntijan arvion perusteella. Ennakointimallien tapauksessa tulisi aina tutkia kalibraatiota ja erottelua, kuten luvussa 5.4 todettiin. Sosiaali- ja terveysalalla yleisista ennakointimallityypeistä riskityökalujen tapauksessa erityisen tärkeää on erottelu ja kuormituksen ennakointityökalujen tapauksessa kalibraatio. Mittareille tulee myös määrittää hyväksyttävät arvot. Määrittämisessä voidaan hyödyntää kirjallisuudesta löytyviä arvoja, joita on kirjattu luvussa 3 tai muuten käytetty vastaavanlaisille malleille. Toinen vaihtoehto on pyrkiä sovelluskohtaisesti määrittämään arvoja päätetyistä tavoitteista lähtien. Mitä useammilla mittareilla mallia arvioidaan, sitä monipuolisempi kuva saadaan ennustekyvystä. Toisaalta tällöin tulosten tulkinta voi olla monimutkaisempaa.

Tavoitteiden määrittelyyn kuuluu myös sen kiteyttäminen, millaisia hyötyjä mallilla toivotaan saavutettavan. Perustavanlaatuisena tavoitteena voi olla esimerkiksi kulujen karsiminen, nopeammat hoitoajat tai sairauden ennaltaehkäisy. Tavoiteltavat hyödyt voivat olla erilaisia eri sidosryhmien näkökulmasta, ja niistä tulisi valita

mallin kannalta olennaisimmat esimerkiksi hyödyntäen kolmijakoa potilas-yksikkö-hoitoketju. Jos mahdollista, hyödyille määritetään tavoitetaso euroina. Vaihetta voi visualisoida piirtämällä päätöspuun (luku 5.1.1), jonka yhtenä haarana on nykytila ja toisena tilanne mallin kanssa. Päätöspuuhun saa kirjattua todennäköisyyksiä ja odotettuja hyötyjä eri lopputulemille, joten se myös auttaa hyötylaskelmien teossa. Toiseksi keinoksi hyödyn arviointiin viitekehyksessä suositellaan hyödyn tarkastelua sovelluskohtaisesti määriteltävillä konkreettisilla, sovellusalueen asiantuntijoiden näkökulmasta relevanteilla ja helposti ymmärrettävillä mittareilla. Valittavien mittareiden tulee olla kvantifioitavissa ja laskettavissa olemassa olevasta datasta. Tavoitteiden, hyötyjen ja kustannuksien määrittelystä kerrotaan enemmän luvussa 5.1.2.

**1.E: Päätös jatkosta**

Tavoitteiden määrittämisen jälkeen tehdään päätös siitä, miten mallin kehitystyötä jatketaan. Tilanteesta riippuen voidaan jatkaa mallin kehitystyötä suunnitelmien mukaan, muokata suunnitelmia tai keskeyttää kehitystyö. Seuraavia osa-alueita tulee arvioida päätöstä tehdessä [58, 59].

Mallin toteuttaminen:

- *Datan saatavuus.* Mallin toteuttaminen on sitä helpompaa, mitä vähemmän dataa täytyy yhdistellä eri lähteistä. Otoskoon täytyy olla riittävän suuri ja aikasarjan riittävän pitkä, jotta hyvä malli on mahdollista tehdä.

- *Intuitiivinen potentiaali.* Jos sovellusalueen asiantuntijat arvioivat vastemuuttujan ja potentiaalisten syötemuuttujien välillä olevan relevantti yhteys, on hyvän mallin rakentaminen todennäköisempää.

Mallin vaikuttavuus:

- *Hyötyjen merkittävyys.* Jos malli ratkaisee merkittävää ongelmaa, jo pienelläkin kehityksellä voidaan saada aikaan merkittäviä hyötyjä.

- *Vaikutus toimintaan.* Jotta mallin rakentaminen olisi järkevää, tulisi onnistuneen mallin antamalla tiedolla olla aito vaikutus toimintaan niin, että tavoitellut hyödyt on mahdollista saavuttaa.

- *Integroitavuus käytössä oleviin järjestelmiin.* Jos malli voidaan integroida käytössä oleviin järjestelmiin, se todennäköisemmin siirtyy osaksi normaalia toimintaa ja muuttaa sitä.

- *Mallin skaalautuvuus.* Vaikuttavinta on kehitystyö, jossa toteutettavaa mallia voidaan hyödyntää useammassa kohteessa. Tällöin useilla toimijoilla on samankaltainen tarve, dataa ja järjestelmät. Skaalautuvat ratkaisut tuovat myös liiketoiminnallisia hyötyjä.

## 6.2 Vaihe 5: Mallin arviointi

Jos vaiheessa 1.E on päätetty jatkaa mallin kehitystyötä, toteutetaan malli kehitysprosessin vaiheiden 2-4 mukaisesti (kuva 21). Kehitetyn mallin arviointi kannattaa aloittaa yleiskuvan hankkimisella vertailemalla mallin ennusteita havaintoihin ja vertailumallin tuloksiin esimerkiksi laskemalla keskiarvoja ja piirtämällä niistä erilaisia kaavioita. Mallin arvioinnissa on useita laskennallisia osioita, jotka perustuvat kirjallisuuskatsaukseen (luvut 2-5). Valittujen ennustekyvyn ja hyödyn mittareiden arvot määritetään, siirrettävyytä tutkitaan ja mallin potentiaali otoskoon suhteen selvitetään. Myös mallin käytön aiheuttamat hyödyt ja riskit lasketaan tai arvioidaan. Lopulta tehdään kokonaisarvio, jonka perusteella päätetään jatkosta.

### 5.A: Mittareiden arvojen määrittäminen

Valittujen ennustekyvyn mittareiden arvot määritetään vertailumallille ja kehitetylle mallille. Arvot lasketaan opetusjoukossa sekä sisäisellä validoinnilla (luku 4.2). Sisäisen validoinnin voi toteuttaa esimerkiksi tasapainotetulla 10-ositetulla ristiinvalidoinnilla (luku 4.2.2), jonka tuottamien estimaattien varianssi ja harhaisuus ovat kohtuullisia. Paras menetelmä kuitenkin riippuu otosjoukon ominaisuuksista. Sisäisellä validoinnilla tutkitaan toistettavuutta, mikä on tärkeää, sillä mallia käytetään eri datalla kuin sitä on opetettu, vaikka data olisikin samasta populaatiosta (luku 4.1). Jos mallin ennustekyky on huomattavasti korkeampi opetusjoukossa kuin ristiinvalidoituna, on kyse luultavasti ylisovittumisesta.

Konkreettisilla hyötyyn liittyvillä mittareilla arvioidaan mallin ja vertailumallin hyötyä. Niiden täydellinen laskeminen ei välttämättä onnistu kaikissa tilanteissa, mutta suuntaa-antavan arvion voi muodostaa.

### 5.B: Siirrettävyyden tutkiminen

Siirrettävyyttä kannattaa arvioida, vaikka suunnitelmissa ei olisikaan käyttää mallia toisessa populaatiossa, kuten luvussa 4.3 todetaan. Siirrettävyydeltään hyvä malli toimii luotettavammin myös alkuperäisessä populaatiossa, vaikka data hieman muuttuisikin. Mallin siirrettävyyttä tutkitaan joko toisella otosjoukolla tai jakamalla otos epäsatunnaisesti osiin esimerkiksi ajan suhteen. Malli opetetaan toisella näistä joukoista ja ennustekyvyn ja hyödyn mittareiden arvot lasketaan toisella. Siirrettävyyttä tutkiessa tulee varmistaa sekä opetus- että validointijoukon kokojen olevan riittävän suuria. Luvussa 4.3 kuvataan erilaisia siirrettävyyden näkökulmia ja kumulatiivinen menetelmä niiden arviointiin. Sosiaali- ja terveysalalla olennaisimmat siirrettävyyden näkökulmat ovat ajallinen, maantieteellinen, metodologinen ja vastemuuttujan kirjo. Mitä laajemmin siirrettävyyttä tutkii, sitä paremmin tiedetään, voiko mallin yleistymiseen luottaa.

Mallit opetetaan toimimaan historiadatalla, joten niiden tulee olla ajallisesti siirrettävissä toimiakseen nykyhetkessä ja tulevaisuudessa. Ajallista siirrettävyyttä voi tutkia toteuttamalla malli jonkin ajanjakson käsittävälle datalle ja validoimalla se toisella. Kehitettävä malli voidaan haluta siirtää esimerkiksi toiseen sairaanhoitopiiriin, jolloin maantieteellinen, metologinen ja vastemuuttujan kirjoon liittyvä siirrettävyys

täytyy tutkia. Maantieteellistä siirrettävyyttä voidaan tutkia hyödyntämällä aineistoa toiselta alueelta. Tämä tarkoittaisi esimerkiksi samankaltaista dataa useasta sairaanhoitopiiristä. Koska käytännöt tietojen kirjaamiseen usein vaihtelevat eri toimipisteissä, saadaan usein vahvistusta metodologiseen siirrettävyyteen. Alueittain myös populaation ominaisuudet vaihtelevat, jolloin vastemuuttujan kirjoon liittyvä siirrettävyys tulee usein tutkituksi. Jos yleistyvyydestä halutaan erityisen luotettavaa kuvaa, kannattaa tutkia myös siirrettävyyttä eri kontekstiin, eli esimerkiksi validoida perusterveydenhuollon datalla erikoissairaanhoidon aineistolla kehitettyä mallia. Tällöin voidaan tutkia siirrettävyyttä populaatioon, jonka ominaisuudet eroavat merkittävästi alkuperäisestä.

### 5.C: Mallin potentiaali määrittäminen aineiston koon suhteen

Mallin potentiaalin arvioimisessa hyödynnetään sitä, että mallin opettamiseen käytetyn aineiston koon kasvaessa mallin ennustekyky yleensä kasvaa asymptoottisesti kohti kynnysarvoa. Tämän käyttäytymisen avulla voidaan arvioida mallin kokonaispotentiaalia, ja päätellä, kuinka lähellä sitä ollaan. Toisaalta voidaan arvioida, kuinka paljon lisää dataa tarvittaisiin, jotta ennustekyky kasvaisi halutuksi. Menetelmä tämän arvioimiseen on *oppimiskäyrä*, jonka käyttö on selitetty luvussa 4.4.1. Sen perusajatuksena on laskea ennustekyky usealla eri opetusjoukon koolla ja sovittaa näin saatuihin pisteisiin potenssilakia noudattava käyrä. Jos oppimiskäyrän perusteella vaikuttaa siltä, että ennustekyvyn kynnysarvoa ei ole vielä saavutettu, kannattaa harkita mallin toteuttamista lisädatalla. Toisaalta, jos kynnysarvo on saavutettu, ei mallia voi enää parantaa lisää dataa hankkimalla.

### 5.D: Mallin käytön hyötyjen ja riskien arviointi

| Taso 0 | Mallin käytön riskit | | | | | |
|---|---|---|---|---|---|---|
| Taso 1 | Potilas | | Yksikkö | | Hoitoketju | |
| Taso 2 | Riski 1 | Riski 2 | Riski 3 | Riski 4 | Riski 5 | Riski 6 |

Kuva 22: Riskienosituksessa riskit kuvataan hierarkkisesti edeten kohti yksityiskohtaisempia riskejä. [62]

Mallin käytön keskeiset hyödyt ja riskit arvioidaan käyttäen kolmijakoa potilas-yksikkö-hoitoketju. Riskien tunnistamisessa ja niiden merkittävyyden arvioinnissa voi hyödyntää esimerkiksi riskienositusta (*risk breakdown structure*, RBS), jossa riskejä etsitään hierakkisesti edeten aina yksityiskohtaisempiin riskeihin [62]. Kuvassa 22 on esimerkki riskienosituksesta, jossa taso 1 kuvaa kolmijakoa potilas-yksikkö-hoitoketju. Kaavion muoto ja tasojen lukumäärä vaihtelee tilanteen mukaan. Riskien suuruutta voi määrittää esimerkiksi euroina. Vastaavasti voi määrittää myös mallin käytön hyötyjä, joita myös verrataan vaiheessa yksi arvioituihin odotettuihin hyötyihin.

### 5.E: Kokonaisarvio

Kokonaisarviota varten analyysit kootaan yhteen ja niitä verrataan määriteltyihin tavoitteisiin. Malli myös annetaan mallintamisen ja sovellusalueen asiantuntijoille arvioitavaksi. Mallin tulisi asiantuntijanäkemyksen mukaan olla toimiva ja hyödyllinen sekä käyttöönotettuna vaikuttaa toimintaan tavoitteiden mukaisesti. Analysoitavat asiat on listattu alle:

Ennustekyvyn ja hyödyn mittareiden arvot:

- Tulokset vastaavat määritettyjä tavoitteita
- Malli antaa parempia tuloksia kuin vertailumalli
- Yleistyvyys on riittävällä tasolla

Mallin potentiaali:

- Malli on saavuttanut potentiaalinsa aineiston koon puolesta

Hyödyt ja riskit

- Mallilla voidaan saavuttaa riittävät hyödyt
- Mallin käytön riskit ovat kohtuulliset

Asiantuntijoiden arvio:

- Asiantuntijoiden mielestä malli vaikuttaa toimivalta
- Mallin tuloksilla voidaan vaikuttaa toimintaan hyödyllisesti

Lähtökohtaisesti kaikkien muistilistan kohtien tulisi toteutua. Saattaa kuitenkin olla, että esimerkiksi kaikki mittarit eivät saavuta tavoitettaan, jolloin vaaditaan asiantuntijoiden arvioita tavoitteiden keskinäisestä tärkeydestä ja riittävästä tasosta. Analyysiä tehdessä tulee pitää mielessä, mikä on kyseisen sovelluskohteen kannalta olennaista.

Analyyseihin pohjautuen tehdään *päätös* mallin tulevaisuudesta. Päätösvaihtoehtoja on kolme:

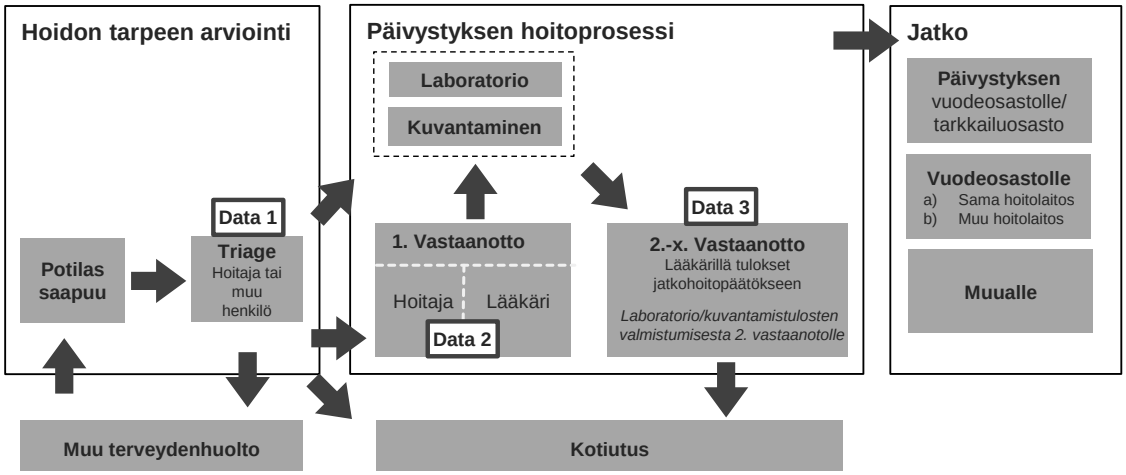
- Malli on riittävän hyvä ja se voidaan ottaa käyttöön. Siirrytään ennakointimallien kehittämisen prosessissa (kuva 21) vaiheeseen 6, eli mallin integroimiseen.
- Mallia tulee kehittää edelleen. Siirrytään tilanteesta riippuen johonkin vaiheista 1-4.
- Lopetetaan mallin kehittäminen. Tilanteessa, jossa malli ei ole saavuttanut tavoitteitaan eikä tavoitteiden saavuttaminen nykytilanteessa ole todennäköistä vaikka mallia kehitettäisiin, kehitystyö lopetetaan.

# 7 Viitekehyksen soveltaminen päivystystoimintaan liittyvään ennakointimalliin

Luvussa 6 määritettiin viitekehys sosiaali- ja terveysalan ennakointimallien arviointiin laajentamalla ennakointimallien kehittämisen prosessia. Tässä luvussa viitekehystä sovelletaan päivystystoimintaan liittyvään ennakointimalliin. Soveltaminen toteutettiin ensisijaisesti viitekehyksen toimivuuden arvioimiseksi, eikä niinkään uuden mallin kehittämiseksi. Viitekehyksessä suositellut asiat käytiin läpi joko haastattelemalla mallia suunnittelemassa ja toteuttamassa ollutta sovellusalueen asiantuntijaa (haastattelu [61]) tai tekemällä analyysit viitekehyksen ohjeistuksen mukaisesti. Haastattelussa käytetty kysymysrunko on liitteessä A. Luvussa 7.1 käsitellään ennakointimallin kehittämisen prosessin vaiheeseen 1 liittyvät kysymykset ja luvussa 7.2 vaiheeseen 5, eli mallin arviointiin liittyvät kysymykset.

## 7.1 Vaihe 1: Ilmiön ymmärtäminen, ongelmien kartoitus ja tavoitteiden määrittäminen

Ennakointimallin kehittämisen prosessin vaiheen 1 (luku 6.1) mukaisesti muodostettiin ymmärrys päivystyksestä ilmiönä. Ilmiöön liittyviä ongelmia kartoitettiin ja ennakointimallilla ratkaistavaksi ongelmaksi valittiin päivystyksen ruuhkautuminen. Tämän jälkeen tehtiin alustava suunnitelma mallista ja määritettiin sille tavoitteet.




Kuva 23: Päivystyksen prosessikaaviossa on kuvattuna keskeiset toimijat, potilasvirrat ja informaation saatavuus [63]

### 1.A: Ilmiön ymmärtäminen

Osiossa tavoitteena on ymmärtää käsiteltävä ilmiö syvällisesti. Päivystyksestä toteutettu yksinkertaistettu prosessikaavio löytyy kuvasta 23. Potilaan saapuessa tehdään hoidontarpeen arviointi (triage), jonka pohjalta potilas kotiutetaan tai ohjataan muualle terveydenhuoltoon, kuvantamis- tai laboratoriotutkimuksiin tai lääkärin tai hoita-

jan vastaanotolle. Vastaanotolla potilas tutkitaan, jonka jälkeen hänet kotiutetaan tai ohjataan kuvantamis- tai laboratoriotutkimuksiin tai jatkohoitoon. Kuvantamis- tai laboratoriotutkimusten tulosten valmistuttua toisella vastaanottokäynnillä tehdään taas päätös jatkosta, eli kotiutetaanko potilas, tilataanko lisätutkimuksia vai ohjataanko potilas jatkohoitoon. Jatkohoito voi olla esimerkiksi päivystyksen tarkkailu- tai vuodeosastolla tai vuodeosastolla samassa tai muussa hoitolaitoksessa. Hoidontarpeen arviointivaiheessa tiedossa on potilaan ja käynnin perustiedot sekä edellisten käyntien tiedot (Data 1). Ensimmäisen vastaanottokäynnin jälkeen tiedetään lisäksi, mitä lisätutkimuksia on tilattu (Data 2). Toisen vastaanottokäynnin kohdalla tiedetään myös lisätutkimusten tulokset (Data 3). Noin 30% päivystykseen saapuvista potilaista päätyy jatkohoitoon vuodeosastolle. [61]

Prosessikaavioon (kuva 23) on kirjattu keskeiset toimijat, toimenpiteet, potilasvirta ja eri vaiheissa tiedossa oleva informaatio. Siitä selviää myös vaiheet, joissa tehdään päätöksiä. Kuvassa esitetyn kaavion lisäksi päivystyksen ennakointimallin kehitystyön alussa ymmärrettiin myös ilmiöön liittyviä dynamiikkoja ja hoitoketjun kokonaisuus laajemmin. Lisäksi tiedettiin, millaisia muuttujia kirjataan tietojärjestelmään eri kohdissa ja kuka sen tekee. [61]

### 1.B: Ongelmien kartoitus ja valinta

Seuraavaksi perehdytään ilmiöön liittyviin ongelmiin. Päivystyksestä löydetyt ongelmat liittyvät kaikki vahvasti ruuhkautumiseen, joka on kansainvälisestikin merkittävä ongelma [64]. Päivystyksen ruuhkautumisella tarkoitetaan sitä, että tunnistettu päivystyspalveluiden tarve ylittää potilaiden hoitoon tarjolla olevat resurssit joko päivystysosastolla, sairaalassa tai molemmissa [64]. Potilaan kannalta ruuhkautuminen tarkoittaa potilastyytyväisyyden laskua, potilasturvallisuuden heikkenemistä ja ennustettavuuden puutetta. Yksikön kannalta ruuhkautuminen tarkoittaa työntekijöiden stressiä ja hoidon aloittamisen viivästymistä. Lisäksi resursseja kuluu turhaan hoitoa ja poislähtöä odottavien potilaiden hoitoon, mikä kasvattaa resurssitarvetta ja kustannuksia. Hoitoketjun näkökulmasta informaation kulku eri toimijoiden välillä on vaikeaa. [61]

Ruuhkautumiseen vaikuttavat lukuisat asiat, jotka liittyvät potilaiden saapumiseen, läpimenoon ja poistumiseen päivystyksestä [64]. Päivystykseen saapuvien potilaiden määrä vaihtelee ennalta arvaamattomasti ja jos päivystys toimii nopeasti, sinne saatetaan lähettää potilaita, jotka kannattaisi kokonaisuuden kannalta hoitaa muualla. Päivystyksen henkilöresurssit saattavat olla liian alhaiset [64]. Toisaalta jatkohoito-osastoilla ei osata varautua päivystyksestä saapuviin potilaisiin, jolloin heitä ei voida lähettää päivystyksestä eteenpäin. Ongelmista ratkaistavaksi valittiin jatkohoito-osastojen varautuminen päivystyksestä saapuviin potilaisiin, koska se on koko hoitoketjun pullonkaula, joka kuluttaa turhaan päivystyksen resursseja ja hidastaa koko prosessia. Ongelman merkittävyyden lisäksi myös muut viitekehyksessä olevat valintaa tehdessä mietittävät näkökulmat tukivat päätöstä. Ennakointimalli vaikutti olevan järkevä ja tehokas tapa ratkaista kyseistä ongelmaa, hyvää dataa oli tarjolla, onnistunut ennakointimalli voisi vaikuttaa toimintaan ja päivystyksen ruuhkautuminen on toimijoiden tärkeä kehityskohde. [61]

### 1.C: Mallin toteutuksen suunnittelu

Viitekehyksessä mallin toteutuksen suunnittelussa määritetään konkreettisemmin kehitettävän mallin piirteitä. Päivystyksen mallissa ennustettavaksi vastemuuttujaksi valittiin potilaan siirtyminen sairaalan vuodeosastolle. Mallissa päivystyksessä oleville potilaille lasketaan todennäköisyys siirtyä osastolle, kynnysarvon perusteella määritetään potilaille binäärinen ennuste siirtymisestä ja ennusteet summataan yhteen ennustetuksi kokonaismääräksi osastolle siirtyviä. Mallin konkreettisena tavoitteena on helpottaa vuodeosaston valmistautumista päivystyksestä tuleviin potilaisiin. Riittävän ajoissa saatavan ennusteen avulla heidän on mahdollista lähettää potilaita kotiin tai muille osastoille ja niin tehdä tilaa saapuville potilaille. Potentiaalisimmiksi vastemuuttujiksi arvioitiin tulosyy, tulotapa, tehty kiireellisyysarvio, ikä, tilatut laboratorio- ja kuvantamistutkimukset sekä potilaan käyntihistoria. Myöhemmin selvisi, ettei kaikkia muuttujia ollut saatavilla. Malli toteutettaisiin integroimalla se asiakkaalla käytössä olevaan tietojärjestelmään, jossa olevien tietojen avulla se laskee ennusteen ja esittää sen visuaalisesti. [61]

### 1.D: Tavoitteiden määrittely

Mallin ennustekyvyssä olennaisinta on saada määritettyä keskimäärin oikea tulos, jotta summattuna ennuste on oikeaa kokoluokkaa. Vuodeosastoilla ei ollut käytössä eksplisiittistä mallia saapuvien potilaiden määrän arviointiin, vaikka implisiittisesti saapuvien potilaiden määrää jonkin verran arvioidaan esimerkiksi vuorokaudenajan perusteella [61]. Vertailumalliksi valittiin päivystyksestä vuodeosastolle saapuvien potilaiden tuntikeskiarvo. Tavoitteeksi valittiin, että kehitettävä malli toimii paremmin kuin tämä vertailumalli. Koska mallin lopullinen tavoite on ennustaa, kuinka monta potilasta osastolle saapuu, valittiin hyötymittareiksi keskimääräinen ja suurin virhe, jota ennuste tekee suhteessa havaintoon. Virhe mitataan yksikössä potilasta osastolle/tunti. Erityisen haitallista päivystysosastolle on suuret resurssivajaukset, jolloin mallin olisi tärkeää pystyä vähentämään niitä. Tämän takia päätettiin tarkastella myös sitä, kuinka suuren osan aikaa mallin virhe on pienempi kuin jokin kynnysarvo. Kynnysarvoa vaihdellaan välillä 1-5 potilasta.

Koska lopullinen malli on yksilötason sijaan summa yksilöiden yli, eivät ennustekyvyn mittarit suoraan kuvaa lopullisen mallin ennustekykyä. Ennustekyvyn mittareista päätettiin kuitenkin tarkastella luokittelun tarkkuutta ja AUC-arvoa (luku 3.2). Luokittelun tarkkuuden tavoitearvoksi asetettiin 0.75, jotta se ylittää luokittelun tarkkuuden triviaalilla mallilla, joka ennustaa kaikkien potilaiden päätyvän muualle kuin vuodeosastolle. AUC tavoitearvoksi puolestaan asetettiin kirjallisuudessa yleisesti hyväksi malliksi tulkittava raja 0.75 (luku 3.2.7).

Tässä vaiheessa myös pyrittiin arvioimaan onnistuneen mallin tuottamia hyötyjä. Ilmiön monimutkaisuuden ja takaisinkytkentöjen takia tämä todettiin vaikeaksi. Kuitenkin, koska päivystyksen ruuhkautuminen on kansainvälisestikin ajatellen merkittävä ongelma, uskottiin että mallin tuottama pienikin parannus toimintaan voi tuottaa merkittäviä hyötyjä kokonaisuudelle.

**1.E: Päätös jatkosta**

Ennakointimallin kehittämisen prosessin ensimmäisen vaiheen lopussa tehdään päätös siitä, kannattaako kehitystyötä jatkaa. Päätöstä tehdessä suunniteltua mallia arvioidaan sen toteuttamisen potentiaalin ja vaikuttavuuden näkökulmista. Toteuttamiseen liittyen datan saatavuus arvioitiin riittäväksi, sillä sitä oli tarjolla yli 80000 yksittäistä päivystyksen käyntiä jatkohoitotietoineen vuoden ajanjaksolta. Lisäksi asiantuntijan arvion mukaan potentiaaliset syötemuuttujat todennäköisesti voisivat ennustaa valittua vastemuuttujaa, vaikka kaikkia vaiheessa 1.C listattuja syötemuuttujia ei ollut saatavilla.

Mallin vaikuttavuuden näkökulmasta ratkaistava ongelma arvioitiin niin merkittäväksi, että myös mallin tuottamat hyödyt ovat riittävän suuria. Ennuste päivystyksestä vuodeosastolle siirtyvien potilaiden määrästä auttaisi osastoja ennakoimaan vuodepaikkojen tarvetta ja lähettämään potilaita tarpeen mukaan eteenpäin, jolloin päivystyksen potilaat voisivat sujuvammin siirtyä jatkohoitoon. Malli olisi myös integroitavissa vuodeosastolla käytössä olevaan järjestelmään. Päivystysosastoja on Suomessa kymmeniä ja niillä on samankaltainen tarve, dataa ja järjestelmät, joten malli olisi hyvin skaalattavissa useaan kohteeseen. Viitekehyksen perusteella kehitystyötä kannattaa jatkaa.

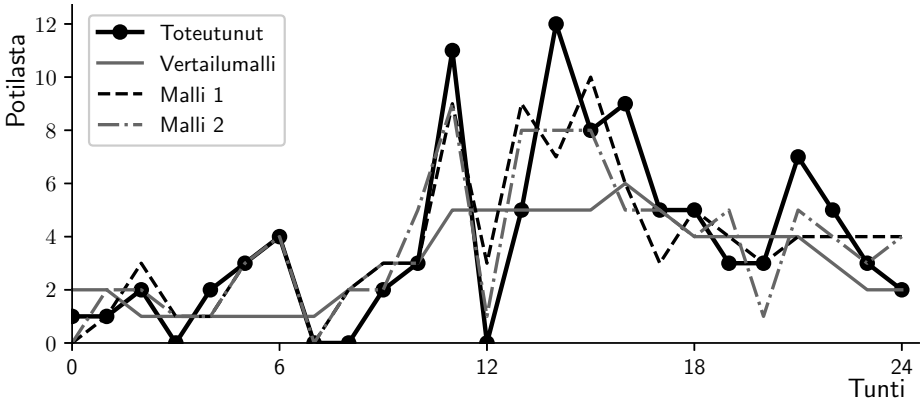
## 7.2 Mallin arviointi

Taulukko 10: Päivystyksen ennakointimallien lopulliset syötemuuttujat

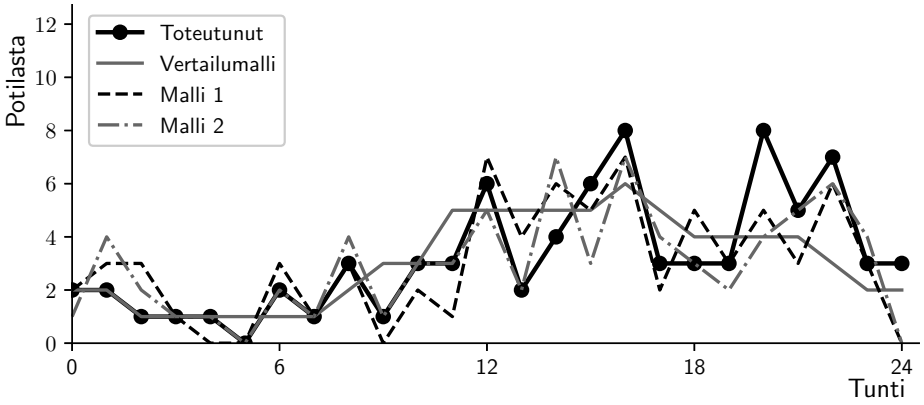
| | Malli 1 | Malli 2 |
|---|---|---|
| Ikä | x | x |
| Sukupuoli | x | x |
| Onko keskuskaupunki | x | x |
| Onko aamuvuoro | x | x |
| Onko iltavuoro | x | x |
| Onko yövuoro | x | x |
| Onko viikonloppu | x | x |
| Triage | x | x |
| Onko laboratoriotilauksia | | x |
| Onko kuvantamistutkimustilauksia | | x |
| Lääkärin vai hoitajan käynti | | x |
| Onko erikoissairaanhoidon käynti | | x |

Ennakointimalli rakennettiin hyödyntäen ennakointimallin kehittämisen prosessin (luku 6) vaiheita 2-4. Lopputuloksena oli kaksi logistisella regressiolla toteutettua mallia, joissa oli eri määrä syötemuuttujia. Ensimmäinen malleista toteutettiin aineistolla, joka on käytössä päivystyksen prosessin (kuva 23) hoidontarpeen arvioinnissa

(data 1) ja toinen laboratorio- ja kuvantamistutkimusten tilaamisen jälkeen (data 2). Lopullisissa malleissa olevat syötemuuttujat on esitelty taulukossa 10. Käytetty aineisto oli anonymisoitu. Opetusjoukkona kummallekin mallille käytetiin kuuden kuukauden pituista aineistoa, jossa oli yhteensä yli 42 000 potilaskäyntiä. Ajallista siirrettävyyttä puolestaan arvioitiin myöhemmin kerätyllä puoli vuotta kattavalla validointijoukolla, jossa oli yhteensä hieman alle 40 000 potilaskäyntiä.




Kuva 24: Toteutuneet, vertailumallin mukaiset sekä kummankin mallin ennustamat potilasmäärät tunneittain yhtenä maanantaina.




Kuva 25: Toteutuneet, vertailumallin mukaiset sekä kummankin mallin ennustamat potilasmäärät tunneittain yhtenä lauantaina.

Kuvissa 24 ja 25 esitetään toteutuneet, vertailumallin ja kummankin ennakointimallin ennustamat tuntikohtaiset potilasmäärät yhden validointijoukkoon kuuluvan esimerkkipäivän ajalta. Kuvassa 24 päivä on maanantai ja kuvassa 25 lauantai. Vertailumalli on aineistosta lasketut potilasmäärän tuntikeskiarvot. Vertailumallissa ei huomioida viikonpäiviä. Kuvista huomataan toteutuneiden potilasmäärien vaihtelevan vuorokauden ajan mukaan ja muuttuvan paljon tunneittainkin. Kehitetyt ennakointimallit näyttäisivät seuraavan vertailumallia paremmin toteutumaa.

Mallien rakentamisen jälkeen toteutettiin ennakointimallien kehittämisen prosessin vaihe viisi eli arviointi (luku 6.2). Valituille mittareille laskettiin arvot ja mallin yleistyvyyttä, potentiaalia, hyötyjä sekä riskejä arvioitiin. Lopulta tehtiin kokonaisarvio.

## 5.AB: Mittareiden arvojen määrittäminen ja siirrettävyyden tutkiminen

Taulukko 11: Sekaannusmatriisi mallille 1

| Ennuste | Havainto | | |
|---|---|---|---|
| | Kyllä | Ei | Summa |
| Kyllä | 7767 | 4963 | 12630 |
| Ei | 5931 | 21263 | 27194 |
| Summa | 13698 | 26126 | 39824 |

Taulukko 12: Sekaannusmatriisi mallille 2

| Ennuste | Havainto | | |
|---|---|---|---|
| | Kyllä | Ei | Summa |
| Kyllä | 8806 | 3812 | 12618 |
| Ei | 4892 | 22314 | 27206 |
| Summa | 13698 | 26126 | 39824 |

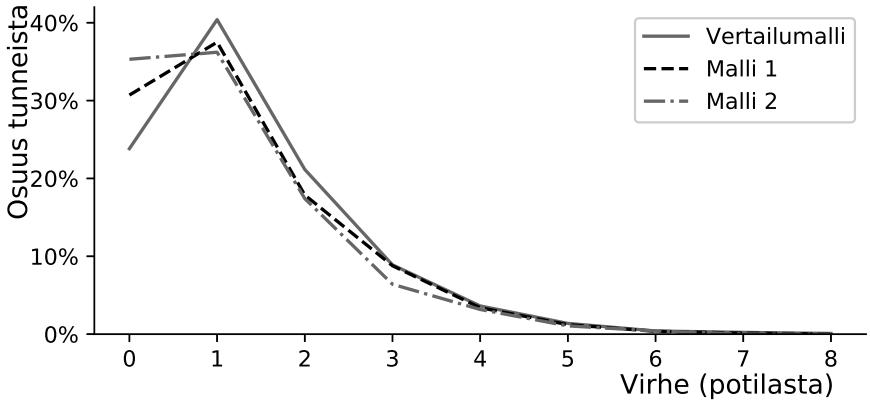
Validaatiojoukossa laskettu sekaannusmatriisi mallille 1 on taulukossa 11 ja mallille 2 taulukossa 12. Kumpikin malli ennustaa hieman havaintoja useammin, että potilas ei siirry jatkohoitoon osastolle. Taulukosta laskettuna oikeiden positiivisten ja negatiivisten määrät mallille 1 on $N_{TP,malli1} = 7767$ ja $N_{TN,malli1} = 21263$ ja mallille 2 on $N_{TP,malli2} = 8806$ ja $N_{TN,malli1} = 22314$. Koska otosjoukko on saman kokoinen kummallakin mallille, on mallilla 2 korkeammat oikeiden positiivisten ja oikeiden negatiivisten osuudet kuin mallilla 1.

Taulukko 13: Päivystyksen ennakointimallin suorituskyvyn mittareiden arvot laskettuna vertailumallille ja kummallekin kehitetylle mallille (malli 1 ja 2). Mittareiden arvot laskettiin opetusjoukossa, ristiinvalidointia käyttäen sekä ajallisesti erillisessä joukossa. Hyötymittarien arvot laskettiin vain ajallisesti erillisessä joukossa.

| Mittari | Tavoite | Vertailumalli | Malli 1 opetusjoukossa | Malli 1 ristiinvalidoituna | Malli 1 ajallinen | Malli 2 opetusjoukossa | Malli 2 ristiinvalidoituna | Malli 2 ajallinen |
|---|---|---|---|---|---|---|---|---|
| Luokittelun tarkkuus | 0.7 | - | 0.74 | 0.74 | 0.73 | 0.79 | 0.79 | 0.78 |
| AUC | 0.75 | - | 0.78 | 0.78 | 0.76 | 0.85 | 0.85 | 0.85 |
| Keskimääräinen virhe (potilas) | <1.4 | 1.4 | - | - | 1.2 | - | - | 1.1 |
| Suurin virhe | <8 | 8 | - | - | 8 | - | - | 7 |
| Virhe 0 potilasta | >24% | 24% | - | - | 30% | - | - | 36% |
| Virhe 0-1 potilasta | >64% | 64% | - | - | 67% | - | - | 73% |
| Virhe 0-2 potilasta | >85% | 85% | - | - | 86% | - | - | 90% |
| Virhe 0-3 potilasta | >90% | 90% | - | - | 95% | - | - | 96% |
| Virhe 0-4 potilasta | >95% | 95% | - | - | 98% | - | - | 99% |

Arvot valituille ennustekyvyn ja hyödyn mittareille laskettiin luvussa 6.2 ohjeistetun mukaisesti (taulukko 13). Mittareiden arvot laskettiin vertailumallille sekä kehitetyille malleille kahdella tavalla. Ensin ne laskettiin opetusjoukossa, eli käyttämällä samaa aineistoa mallin opettamiseen ja arviointiin. Sitten toteutettiin sisäinen validaatio, eli data jaettiin useita kertoja opetus- ja validointijoukkoihin käyttäen menetelmänä tasapainotettua 10-osioitua ristiinvalidointia. Koska vertailumalli on tuntikeskiarvo, niin sille ei voida määrittää käyntitason ennustekyvyn mittareita, kuten luokittelun tarkkuus ja AUC. Opetusjoukossa laskettuna ennustekyvyn mittarit vastaavat ristiinvalidoinnilla saatuja tuloksia, joten mallit eivät vaikuta olevan ylisovittuneita.

Ajallisesti erillisessä joukossa arvot ovat hieman matalampia. Mallit kuitenkin toimivat hyvin myös kyseisessä joukossa, eli mallin ajallinen siirrettävyys on riittävällä tasolla. Datan puolesta ei ollut mahdollista tutkia siirrettävyyttä maantieteellisellä tai domain validaatiolla.



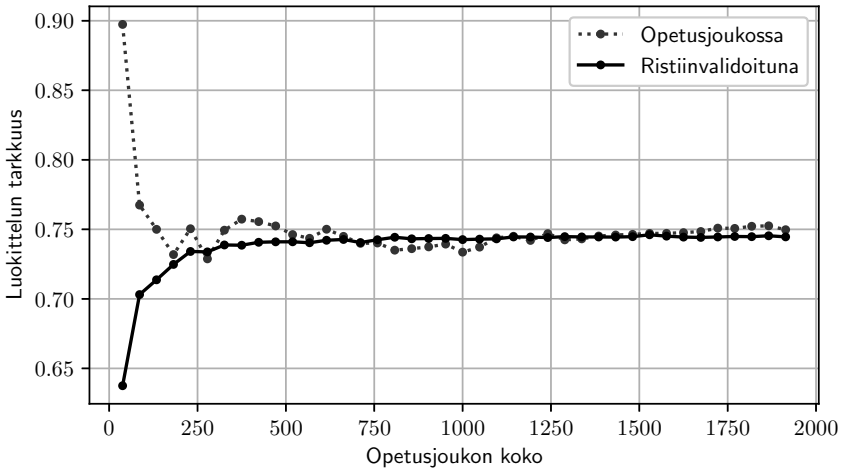

Kuva 26: Eri suuruisten virheiden prosentuaaliset osuudet vertailumallille sekä kummallekin kehitetylle mallille

Eri suuruisten virheiden prosentuaaliset osuudet laskettuna ajallisesti erillisessä joukossa on esitetty kuvassa 26. Kuvasta huomataan, että kaikilla malleilla suurin osuus on virheen ollessa 1 ja osuus pienenee virheen koon kasvaessa. Virheettömien tuntien osuus on suurin mallilla 2 ja pienin vertailumallilla. Virheellisten tuntien osuus on vastaavasti suurin vertailumallilla ja pienin mallilla 2 riippumatta virheen suuruudesta.

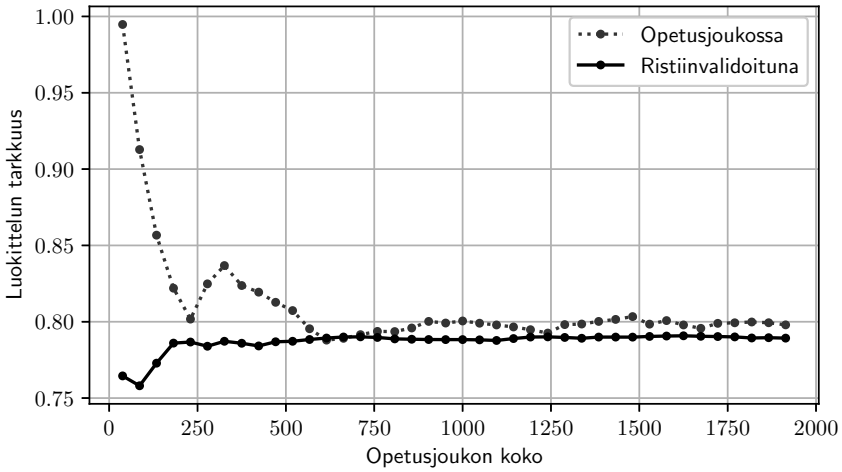
Kaikkien hyötymittarien näkökulmasta molemmat mallit täyttävät tavoitteen, sillä ne toimivat paremmin kuin vertailumalli. Kummallakin mallilla keskimääräinen virhe on merkittävästi pienempi kuin vertailumallilla. Suurimmassa virheessä eroa ei ole merkittävästi. Pienten virheiden osuutta kartoittavilla mittareilla erityisesti malli 2 tuottaa merkittävän parannuksen verrattuna vertailumalliin. Esimerkiksi mittarilla 'Virhe 0-2 potilasta' malli 2 tuottaa 5 prosenttiyksikön kasvun vertailumalliin verratuna, mikä tarkoittaa vuoden aikana yli 18 päivää. Eli yli 18 päivänä vähemmän mallin 2 ennuste eroaa havainnoista kolmella tai useammalla potilaalla verrattuna vertailumalliin.

### 5.C: Mallin potentiaalin määrittäminen aineiston koon suhteen

Viitekehyksessä mallin potentiaalia arvioidaan oppimiskäyrän avulla. Oppimiskäyrässä ennustekyvyn mittarina käytettiin luokittelun tarkkuutta, jolle laskettiin arvo sekä opetusjoukossa että ristiinvalidoinnin avulla. Tulokset kummallekin arvioidulle mallille löytyvät kuvista 27 ja **??**. Kuvista huomataan, että oppimiskäyrä tasaantuu jo opetusjoukon koon ollessa joitakin satoja eikä näytteiden määrän lisääminen paranna ennustekykyä. Opetukseen käytettävän aineiston koon ollessa yli 40000, voidaan todeta mallien saavuttaneen potentiaalinsa näytekoon puolesta. Koska mallin ennustekyky ei vähene vaikka opetusjoukon kokoa pienennettäisiin, voidaan myös päätellä, että ristiinvalidoinnilla saadut ennustekyvyn arviot tuskin ovat pessimistisiä.




Kuva 27: Oppimiskäyrä päivystyksen mallille 1




Kuva 28: Oppimiskäyrä päivystyksen mallille 2

### 5.D: Mallin käytön hyötyjen ja riskien arviointi

Lisäksi arvioidaan mallin hyötyjä ja riskejä. Tarkkojen mallin tuottamien hyötyjen laskeminen todettiin vaikeaksi, mutta päivystyksen ruuhkautumisen aiheuttaman ongelman suuruuden takia voidaan olettaa mallilla saavutettavien hyötyjen olevan riittäviä. Ennustekyvyn ja hyötymittareiden arvoista voidaan huomata mallin selkeästi helpottavan vuodeosaston resursointia vertailumalliin verrattuna ja vaiheessa yksi arvioitiin mallin voivan vaikuttaa toimintaan. Mallin ansiosta vuodeosastolle saapuvista potilaista saadaan ennuste jo useita tunteja ennen heidän saapumistaan, jolloin osaston henkilökunnalla on aikaa varautua siihen.

Mallin käytöllä ei koettu olevan merkittäviä riskejä. Vuodeosastolla saatetaan kokea kovempaa painetta lähettää hoitoa tarvitseviakin potilaita kotiin, jos malli ennustaa paljon saapuvia potilaita. Päätökset tekevät kuitenkin aina asiantuntijat, joten mallin tuottama riski potilasturvallisuudelle ei ole merkittävä. Malli toimii päätöksenteon tukena eikä sen avulla suoraan toteuteta resursointia, joten mallin ennusteita myös arvioidaan kriittisesti, jolloin isommat virheet ennusteissa luultavasti huomataan ja voidaan selvittää.

### 5.E: Kokonaisarvio

Viitekehyksessä esitellyssä kokonaisarviossa kootaan yhteen analyysit ennustekyvyn ja hyödyn mittareista sekä mallin potentiaalista, hyödyistä ja riskeistä. Lisäksi asiantuntijat arvioivat mallin. Tämän jälkeen tehdään päätös jatkosta.

Ennustekyvyn mittareiden suhteen molemmat mallit saavuttavat tavoitteet mutta malli 1 niukasti. Molemmat mallit antavat hyötymittareiden suhteen parempia tuloksia kuin vertailumalli eli täyttävät siltä osin tavoitteensa. Yleistyvyyttä tutkittiin vain ajan suhteen, mutta tässä sovelluskohteessa se arvioitiin riittäväksi. Malli 2 antoi kaikista osa-alueista hieman parempia tuloksia kuin malli 1, mutta toisaalta sen tulokset saadaan myöhemmin kuin mallin 1. Mallin 2 hyöty käytännössä saattaa siis olla rajoitetumpi, sillä vuodeosastoille ei jää yhtä paljon aikaa varautua saapuviin potilaisiin.

Oppimiskäyrän analysoinnilla todettiin mallien saavuttaneen potentiaalinsa aineiston koon suhteen. Mallien tuottamien hyötyjen arvioitiin olevan riittävät ja riskit arvioitiin mataliksi. Lisäksi asiantuntijat arvioivat mallit toimiviksi ja niiden tuottamien ennusteiden vaikuttavan toimintaan hyödyllisesti. Kaikkien viitekehyksen kokonaisarviossa olevien kohtien mukaan molemmat mallit ovat siis riittävän hyviä käyttöönotettavaksi, jolloin seuraavana vaiheena ennakointimallien toteuttamisen prosessissa on mallin integrointi tuotantojärjestelmiin.

# 8 Johtopäätökset

Diplomityössä toteutettiin laaja katsaus olemassa oleviin ennakointimallien suorituskyvyn arviointimenetelmiin ja muotoiltiin viitekehys siitä, kuinka arviointia tulisi tehdä. Vaikka viitekehyksen sisältö muuttuikin useita kertoja työn aikana, diplomityöprojektin alussa asetetut tavoitteet saavutettiin. Kirjallisuusosiossa esitellään kattavasti menetelmiä niin mallien ennustekyvyn, yleistyvyyden kuin niiden käytön hyötyjen arvioimiseen. Kirjallisuudesta ei löydetty laajasti käytössä olevaa ja yleisesti hyväksyttyä tapaa arvioida kokonaisvaltaisesti mallien suorituskykyä, joten viitekehys on koostettu eri näkökulmista muodostetun kokonaiskuvan pohjalta.

Viitekehyksen avulla arviointinäkökulma tulee osaksi ennakointimallin rakentamisen prosessia suunnittelusta lähtien. Viitekehys ottaa mallin arvioinnin huomioon monipuolisesti, mutta on riittävän kevyt toteutettavaksi käytännön projekteissa. Viitekehyksen käyttö vaatii tilanteesta riippuvaa soveltamista, mikä on sekä hyvä että huono asia. Se on riittävän yleinen, jotta se sopii monenlaisiin projekteihin ja sovelluskohteisiin. Toisaalta käyttö vaatii vahvaa asiantuntemusta niin sovelluskohteesta kuin ennakoivasta analytiikasta, jotta tulokset ovat luotettavia. Kun viitekehystä käytetään erilaisten ennakointimallien rakentamiseen, kannattaa siihen lisätä suosituksia ja hyviä käytänteitä. Ajan kanssa viitekehys muuttuu monipuolisemmaksi ja yksityiskohtaisemmaksi ohjeistukseksi.

Viitekehystä varten pyrittiin määrittämään yleispätevä keino, esimerkiksi jonkinlainen päätöspuu, jolla voitaisiin valita tärkein tai tärkeimmät ennustekyvyn mittarit perustuen mallin ominaisuuksiin tai tavoitteisiin kuten vastemuuttujan tyyppiin tai siihen mikä ennustekyvyssä on tärkeintä. Näkemys parhaista mittareista vaihtelee kuitenkin paljon niin tieteenalojen välillä kuin niiden sisällä eikä kirjallisuudesta tai asiantuntijahaastatteluilla saatu riittävästi tukea niiden valintaan. Tämän takia viitekehyksessä suositellaan, että arvioinnissa käytetään useita erilaisia mittareita, jotta ennustekyvystä saadaan monipuolinen kuva. Lisätutkimuksen avulla voi olla mahdollista valita parhaat mittarit erilaisille malleille esimerkiksi tutkimalla tarkemmin riskityökaluja ja kuormituksen ennakointityökaluja ja muodostamalla niille ohjeistukset ennustekyvyn mittareiden valintaan.

Harvalle ennustekyvyn mittarille löytyy yleisiä arvoja, jolloin mallin voisi todeta olevan ennustekyvyltään riittävä. Se, mikä arvo tarkoittaa riittävää ennustekykyä, riippuu sovelluskohteesta. Tavoitteet ennustekyvylle tuleekin määrittää yhdessä sovellusalueen asiantuntijan kanssa. Työn aikana yritettiin löytää keinoja määrittää hyväksyttävät arvot ennustekyvyn mittareille lähtien mallille asetetuista tavoitteista hyödylle, mutta siinä ei onnistuttu. Viitekehyksessä lähtökohdaksi otettiinkin mallin vertaaminen olemassa olevaan tai triviaaliin malliin, jolloin ennustekyvyn mittareille saadaan sovelluskohtainen vertailuarvo.

Hyöty oli harvoin näkökulmana löydetyissä esimerkeissä ennakointimallien arvioinnista. Viitekehystä toteutettaessa hyödyn tarkastelua pidettiin kuitenkin keskeisenä näkökulmana, sillä hyvä ennustekyky ja yleistyvyys ei vielä takaa, että mallilla on tavoitteiden mukaisia vaikutuksia toimintaan. Työn aikana mallien tuottaman hyödyn laskemisen huomattiin olevan hankalaa. Ilmiöt, joihin ennakointimalleja sovelletaan, ovat usein niin monimutkaisia, että yksinkertaisia keinoja hyödyn laskemiseen ei

löydetty. Jotta hyötyjä oikeassa käytössä voitaisiin varmistaa tutkimuksella, tulisi toteuttaa raskaita vaikuttavuustutkimuksia, mikä käytännössä olisi usein liian työlästä. Hyödyn arviointia saattaa olla mahdollista toteuttaa esimerkiksi simulointimalleilla, mutta ennen kuin uusia menetelmiä voi lisätä viitekehyksen osaksi, tulee tehdä lisätutkimusta.

Parhaimmillaan hyödyn tarkastelu tarkoittaisi kvantitatiivisia arvioita esimerkiksi siitä, kuinka paljon resursseja mallin käytöllä voisi säästää tai elinvuosia lisätä. Se olisi arvokasta niin mallin suunnitteluvaiheessa, jolloin päätetään kannattaako malli kehittää, kuin arviointivaiheessa päätettäessä mallin integroinnista. Viitekehyksen kannalta arvokasta lisätutkimusta ovatkin ohjeistukset mallin tuottaman hyödyn arviointiin. Kustannusvaikuttavuusanalyysi on yksi näkökulma, josta tähän voisi olla apua. Aina ei kuitenkaan ole välttämätöntä selvittää hyötyjen tarkkoja arvoja, vaan varmistaa, että hyödyt ovat suurempia kuin kustannukset ja vertailla eri vaihtoehtojen hyötyjä suuruusluokkatasolla. Esimerkiksi ratkaistavaa ongelmaa valitessa suositellaan arvioimaan niiden merkittävyyden suuruusluokkaa eri näkökulmista eikä laskemaan lukuja täysin auki.

Toiseksi menetelmäksi mallin hyödyn arviointiin suositellaan käytettäväksi konkreettisia, aineistosta laskettavissa olevia hyötymittareita. Niillä voidaan tuoda mallin vaikutuksia konkretiaan ilman uuden tiedon hankkimista. Hyvänä puolena menetelmässä on, että mittareiden laskeminen ei vie paljoa resursseja. Toisaalta mittareiden valinta voi olla haastavaa. Niiden tulisi kuvata ilmiöstä koko hoitoketjun kannalta olennaisia kohtia, jotta vältyttäisiin osa-optimoinnilta. Lisätutkimuksella viitekehystä voitaisiin laajentaa vahvemmaksi avuksi hyötymittareiden määrittämiseen. Hyötymittareiden lisäksi mallin tuottamaa hyötyä voisi konkretisoida visualisoimalla mallin tuloksia erilaisilla kaavioilla. Tulosten visualisoimiseen ei perehdytty tässä diplomityössä, mutta sitä suositellaan tutkittavaksi tulevaisuudessa ja lisättäväksi viitekehyksen osaksi.

Viitekehystä sovellettiin päivystyksen ennakointimalliin sen käytettävyyden ja hyödyllisyyden arvioimiseksi. Tämän perustella viitekehys toimii hyvänä muistilistana asioita, jotka tulee ottaa huomioon mallia suunniteltaessa ja arvioitaessa. Viitekehyksen käyttö pakottaa tarkastelemaan ilmiötä ja mallia kokonaisvaltaisesti ja tekemään punnittuja päätöksiä mallin kehittämisen jatkosta. Oppimiskäyräanalyysin perusteella dataa oli käytössä runsaasti. Jos näin on muissakin sovelluskohteissa, on mahdollisuus toteuttaa monimutkaisiakin malleja ilman pelkoa ylisovittumisesta. Toisaalta yleistyvyydestä ei pystytty tutkimaan kuin ajallista näkökulmaa, joten syötemuuttujiltaan vastaavien aineistojen hankkiminen esimerkiksi useilta alueilta olisi mallin arvioinnin kannalta hyödyllistä. Viitekehystä soveltaessa huomattiin, että mallin tuottaman hyödyn arviointi oli hyvin vaikeaa. Lisää tutkimusta tarvittaisiin, jotta sosiaali- ja terveysalan ennakointimallien hyödyn arviointiin saataisiin tukea.

Vaikka tässä diplomityössä keskityttiin arviointiin on mallien rakentamisen muissakin vaiheissa tehtävänä lukuisia päätöksiä. Viitekehyksen voisi toteuttaa esimerkiksi datan keräämisestä ja prosessoinnista, mallin valinnasta ja kehittämisestä sekä ennakointimallien käyttöönotosta. Tarkempia kysymyksiä voisivat olla esimerkiksi toimintatavat pienten otosjoukkojen tapauksessa, miten oppiva algoritmi tulisi valita riippuen tutkittavasta kysymyksestä tai kuinka ennakointimallien tarjoama

informaatio tulisi käytännössä tuoda sosiaali- ja terveysalan toimijoille, jotta se olisi mahdollisimman vaikuttavaa.

Diplomityö tarjosi menetelmiä mallien arviointiin, mutta ei siihen, millä tavoin heikosti toimivaa mallia tulisi kehittää edelleen ja miten arviointi silloin toteutettaisiin. Ennakointimallit saattavat myös vanhentua ajan kanssa, jos esimerkiksi hoitokäytännöt muuttuvat. Kerran kehitetty ja arvioitu malli ei välttämättä toimi enää vuosien päästä. Mielenkiintoinen tutkimuskohde olisikin analysoida miten ja millä aikavälein malleja tulisi uudelleenarvioida ja toisaalta, miten niitä kannattaisi päivittää uudella datalla.

# Kirjallisuusviitteet


[1] Chintan M Bhatt, Nilanjan Dey ja Amira Ashour. *Internet of Things and Big Data Technologies for Next Generation Healthcare*. Springer, 2017.

[2] Riika-Leena Leskelä, Vesa Komssi, Saana Sandström ym. "Paljon sosiaali- ja terveyspalveluja käyttävät asukkaat Oulussa". *Suomen lääkärilehti* 48.68 (2013), s. 3163–3168.

[3] Trevor Hastie, Robert Tibshirani ja Jerome Friedman. *The elements of statistical learning*. Springer, 2008.

[4] Amy C Justice, Kenneth E Covinsky ja Jesse A Berlin. "Assessing the generalizability of prognostic information". *Annals of internal medicine* 130.6 (1999), s. 515–524.

[5] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook ym. "Assessing the performance of prediction models: a framework for some traditional and novel measures". *Epidemiology (Cambridge, Mass.)* 21.1 (2010), s. 128.

[6] Stuart G Baker, Nancy R Cook, Andrew Vickers ym. "Using relative utility curves to evaluate risk prediction". *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172.4 (2009), s. 729–748.

[7] Yrjänä Hynninen, Tutkimuspäällikkö. *Haastattelu 29.11.2017*. Nordic Healthcare group, Vattuniemenranta 2. Haastattelijana Milja Asikainen.

[8] *Nordic Healtcare Group kotisivut*. `http://nhg.fi`. Luettu: 2017-12-01.

[9] Ewout W Steyerberg. *Clinical prediction models: a practical approach to development, validation, and updating*. Springer, 2008.

[10] Scikit-learn developers. *Choosing the right estimator*. Luettu: 4.1.2018. URL: `http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html`.

[11] James Wu ja Stephen Coggeshall. *Foundations of predictive analytics*. CRC Press, 2012.

[12] Min-Ling Zhang ja Zhi-Hua Zhou. "A review on multi-label learning algorithms". *IEEE transactions on knowledge and data engineering* 26.8 (2014), s. 1819–1837.

[13] Mohamed Aly. "Survey on multiclass classification methods". *Neural Netw* 19 (2005), s. 1–9.

[14] Marina Sokolova ja Guy Lapalme. "A systematic analysis of performance measures for classification tasks". *Information Processing & Management* 45.4 (2009), s. 427–437.

[15] David Martinus Johannes Tax. "One-class classification". Tohtorinväitöskirja. Delft University of Technology, 2001.

[16] Sotiris B Kotsiantis, I Zaharakis ja P Pintelas. "Supervised machine learning: A review of classification techniques". *Emerging artificial intelligence applications in computer engineering* 160 (2007), s. 3–24.

[17] Leo Breiman ym. “Statistical modeling: The two cultures (with comments and a rejoinder by the author)”. *Statistical science* 16.3 (2001), s. 199–231.

[18] Yong-ho Lee, Heejung Bang ja Dae Jung Kim. “How to establish clinical prediction models”. *Endocrinology and Metabolism* 31.1 (2016), s. 38–44.

[19] Andrew J Vickers ja Elena B Elkin. “Decision curve analysis: a novel method for evaluating prediction models”. *Medical Decision Making* 26.6 (2006), s. 565–574.

[20] Cort J Willmott. “Some comments on the evaluation of model performance”. *Bulletin of the American Meteorological Society* 63.11 (1982), s. 1309–1313.

[21] Ilkka Mellin. *Sovellettu todennäköisyyslasku: kaavat ja taulukot*. 2002. URL: `http://salserver.org.aalto.fi/vanhat_sivut/Opinnot/Mat-2.090/pdf_varasto/1_painos.pdf`.

[22] Charles Spearman. “The proof and measurement of association between two things”. *The American journal of psychology* 15.1 (1904), s. 72–101.

[23] David B. Montgomery ja Donald G. Morrison. “A Note on Adjusting R2”. *The Journal of Finance* 28.4 (1973), s. 1009–1013.

[24] Cynthia S Crowson, Elizabeth J Atkinson ja Terry M Therneau. “Assessing calibration of prognostic risk scores”. *Statistical methods in medical research* 25.4 (2016), s. 1692–1706.

[25] Corné AM Roelen, Ute Bültmann, Willem van Rhenen ym. “External validation of two prediction models identifying employees at risk of high sickness absence: cohort study with 1-year follow-up”. *BMC Public Health* 13.1 (2013), s. 105.

[26] Frank Harrell, Kerry Lee ja Daniel B. Mark. “Prognostic/Clinical Prediction Models: Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors”. *Statistics in Medicine* 15.4 (1996), s. 361–387.

[27] Grigorios Tsoumakas ja Ioannis Katakis. “Multi-label classification: An overview”. *International Journal of Data Warehousing and Mining* 3.3 (2006), s. 64–74.

[28] Glenn W Brier. “Verification of forecasts expressed in terms of probability”. *Monthly weather review* 78.1 (1950), s. 1–3.

[29] Ewout Steyerberg, Frank Harrell, Gerard Borsboom ym. “Internal validation of predictive models: efficiency of some procedures for logistic regression analysis.” *Journal of Clinical Epidemiology* 54.8 (2001), s. 774–781.

[30] Prabasaj Paul, Michael L. Pennell ja Stanley Lemeshow. “Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets”. *Statistics in Medicine* 32.1 (2013), s. 67–80.

[31] Ewout W Steyerberg ja Yvonne Vergouwe. “Towards better clinical prediction models: seven steps for development and an ABCD for validation”. *European heart journal* 35.29 (2014), s. 1925–1931.

[32] Abdul Ghaaliq Lalkhen ja Anthony McCluskey. “Clinical tests: sensitivity and specificity”. *Continuing Education in Anaesthesia Critical Care & Pain* 8.6 (2008), s. 221–223.

[33] James A Hanley ja Barbara J McNeil. “The meaning and use of the area under a receiver operating characteristic (ROC) curve”. *Radiology* 143.1 (1982), s. 29–36.

[34] David J. Hand ja Robert J. Till. “A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems”. *Machine Learning* 45.2 (2001), s. 171–186.

[35] Karimollah Hajian-Tilaki. “Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation”. *Caspian journal of internal medicine* 4.2 (2013), s. 627.

[36] Laura Schummers, Katherine P. Himes, Lisa M. Bodnar ym. “Predictor characteristics necessary for building a clinically useful risk prediction model: a simulation study”. *BMC Medical Research Methodology* 16.1 (2016), s. 123.

[37] Giovanni Tripepi, Kitty J. Jager, Friedo W. Dekker ym. “Statistical methods for the assessment of prognostic biomarkers (Part I): Discrimination”. *Nephrology Dialysis Transplantation* 25.5 (2010), s. 1399–1401.

[38] Holly Janes, Margaret S Pepe ja Wen Gu. “Assessing the Value of Risk Predictions by Using Risk Stratification Tables”. *Annals of internal medicine* 149.10 (2008), s. 751–760.

[39] Michael J. Pencina, Ralph B. D’ Agostino, Ralph B. D’ Agostino ym. “Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond”. *Statistics in Medicine* 27.2 (2008), s. 157–172.

[40] Sebastian Raschka. *Model evaluation, model selection, and algorithm selection in machine learning: Part II*. 2016. (Viitattu 04. 12. 2017).

[41] Sebastian Raschka. *Model evaluation, model selection, and algorithm selection in machine learning: Part I*. 2016. (Viitattu 11. 10. 2017).

[42] *A study of cross-validation and bootstrap for accuracy estimation and model selection*, s. 1137–1145.

[43] DB Toll, KJM Janssen, Y Vergouwe ym. “Validation, updating and impact of clinical prediction rules: a review”. *Journal of clinical epidemiology* 61.11 (2008), s. 1085–1094.

[44] Mitchell H Katz. “Multivariable analysis: a primer for readers of medical research”. *Annals of internal medicine* 138.8 (2003), s. 644–650.

[45] Karel G M Moons, Patrick Royston, Yvonne Vergouwe ym. “Prognosis and prognostic research: what, why, and how?” *BMJ* 338 (2009), b375.

[46] Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula ym. “Predicting sample size required for classification performance”. *BMC medical informatics and decision making* 12.1 (2012), s. 8–18.

[47] Corinna Cortes, Lawrence D Jackel, Sara A Solla ym. “Learning curves: Asymptotic values and rate of convergence”. *Advances in Neural Information Processing Systems* (1994), s. 327–334.

[48] Sayan Mukherjee, Pablo Tamayo, Simon Rogers ym. “Estimating dataset size requirements for classifying DNA microarray data”. *Journal of computational biology* 10.2 (2003), s. 119–142.

[49] Marthe R Gold. *Cost-effectiveness in health and medicine*. Oxford university press, 1996.

[50] Karel GM Moons, Andre Pascal Kengne, Diederick E Grobbee ym. “Risk prediction models: II. External validation, model updating, and impact assessment”. *Heart* 98.9 (2012), s. 691–698.

[51] Gary S Collins, Joris A de Groot, Susan Dutton ym. “External validation of multivariable prediction models: a systematic review of methodological conduct and reporting”. *BMC medical research methodology* 14.1 (2014), s. 40.

[52] MG Myriam Hunink, Milton C Weinstein, Eve Wittenberg ym. *Decision making in health and medicine: integrating evidence and values*. Cambridge University Press, 2014.

[53] Raúl Poler, Josefa Mula ja Manuel Díaz-Madroñero. “Decision Theory”. Teoksessa: *Operations Research Problems: Statements and Solutions*. Springer London, 2014, s. 205–280.

[54] Ben Van Calster, Andrew J Vickers, Michael J Pencina ym. “Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures”. *Medical Decision Making* 33.4 (2013), s. 490–501.

[55] Paulus Torkki, Kehitysjohtaja ja Tomi Malmström, Liiketoimintajohtaja. *Haastattelu 1.12.2017*. Nordic Healthcare group, Vattuniemenranta 2. Haastattelijana Milja Asikainen.

[56] Jørgen Hilden ja Thomas A Gerds. “A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index”. *Statistics in medicine* 33.19 (2014), s. 3405–3414.

[57] Nordic Healthcare Group. *Description of NHG approach to building analytics models*. Julkaisematon. 6/2017.

[58] Vesa Komssi, Toimitusjohtaja. *Haastattelu 28.11.2017*. Nordic Healthcare group, Vattuniemenranta 2. Haastattelijana Milja Asikainen.

[59] Tommi Kemppainen, Senior Manager. *Haastattelu 4.12.2017*. Nordic Healthcare group, Vattuniemenranta 2. Haastattelijana Milja Asikainen.

[60] Roberta Tassi. *Service Design tools*. Luettu 5.1.2018. URL: `http://www.servicedesigntools.org/`.

[61] Tomi Malmström, Liiketoimintajohtaja. *Haastattelu 19.12.2017*. Nordic Healthcare group, Vattuniemenranta 2. Haastattelijana Milja Asikainen.

[62] David Hillson. “Using a risk breakdown structure in project management”. *Journal of Facilities management* 2.1 (2003), s. 85–97.

[63] Nordic Healthcare Group. *Päivystyksen hoitoprosessi*. Julkaisematon. 12/2017.

[64] Nathan R Hoot ja Dominik Aronsky. “Systematic review of emergency department crowding: causes, effects, and solutions”. *Annals of emergency medicine* 52.2 (2008), s. 126–136.

# A Teemahaastattelujen rungot

Teemahaastattelussa [7] käytetyt kysymykset:

- Millaisia ennakointimalleja NHG:lla toteutetaan?
- Miten NHG:lla olevia malleja voidaan tyypitella?
- Mitkä ovat eri tyyppien keskeiset piirteet? Mikä on niiden ennustekyvyssä olennaisinta?
- Miten mallien tuottamaa hyötyä voitaisiin arvioida?

Teemahaastattelussa [55] käytetyt kysymykset:

- Millaisia asioita ennakointimalliprojektin alussa tulisi ymmärtää tai selvittää, jotta projektin onnistumispotentiaali olisi mahdollisimman korkea?
- Miten tutkittavasta ilmiöstä voidaan muodostaa riittävä kuva? Mitä asioita tulee selvittää?
- Miten projektille voidaan määrittää tavoitteet?

Teemahaastattelussa [61] käytetyt kysymykset:

- Miten päivystys toimii?
- Mitä olennaisia ilmiön osia päivystyksen prosessikuvasta puuttuu? Ymmärrettiinkö niitä projektin alussa?
- Mitä tiedettiin datan saatavuudesta projektin alussa?
- Mitä ongelmia päivystykseen liittyy potilaan, yksikön ja hoitoketjun näkökulmasta?
- Miten arvioisit näiden ongelmien merkittävyyttä?
- Mitä näkökulmia tulee ottaa huomioon kehityskohdetta valitessa?
- Mikä on mallissa vastemuuttujana?
- Mitkä syötemuuttujan ovat potentiaalisimpia?
- Mikä mallin ennustekyvyssä on olennaisinta?
- Mitä mallilla konkreettisesti halutaan muuttaa?
- Miten mallin tuottamaan hyötyä voitaisiin mitata?

Teemahaastattelussa [58] käytetyt kysymykset:

- Miten voidaan arvioida ennakointimallin onnistumispotentiaalia jo kehitystyön alussa?

Teemahaastattelussa [59] käytetyt kysymykset:

- Ennakointimallin kehitystyöhön lähdettäessä olisi hyvä arvioida, onko malli toteutettavissa ja voisiko onnistuessaan myös aidosti vaikuttaa toimintaan, eli olla hyödyllinen. Miten näitä asioita voisi arvioida jo kehitystyön alussa?
- Edellisessä haastattelussa identifioitiin olennaisiksi teemoiksi onnistumispotentiaalin kannalta ongelman merkittävyys, datan lähteiden laatu, vaikuttaako asiantuntijoiden mielestä syöte- ja vastemuuttujien välillä olevan yhteys, voiko mallin tuottaman informaation avulla muuttaa toimintaa ja onko malli integroitavissa käytössä olevaan järjestelmään. Mitä mieltä olet näistä? Mitä listasta puuttuu?
- Miten näita asiota voisi käytännössä selvittää kehitystyön alkuvaiheessa?

# B Viitekehyksen yhteenveto

| | |
|---|---|
| 1 | Ilmiön ymmärtäminen, ongelmien kartoitus ja tavoitteiden määrittäminen |
| | A Keskeiset toimijat ja toimenpiteet, potilas- ja informaatiovirrat, vaikutuspaikat, vuorovaikutukset ja dynamiikat, hoitoketjun kokonaisuus, ilmiöstä kirjattava tai muuten relevantti data |
| | B Ilmiöön liittyvät ongelmat jaottelulla potilas-yksikkö-hoitoketju, niiden merkittävyys sekä ennakointimallilla ratkaistavan ongelman valinta |
| | C Vastemuuttuja, potentiaaliset syötemuuttujat, käyttäjät ja käyttötapa, miten malli muuttaisi toimintaa |
| | D Vertailumalli, mittarit ja tavoitteet ennustekyvylle ja hyödylle |
| | E Päätös jatkosta perustuen mallin toteuttamis- ja vaikuttamispotentiaaliin |
| 2 | Lähtödatan keräys eri lähteistä |
| 3 | Datan esikäsittely, muokkaus ja rajaus |
| 4 | Muuttujien ja mallin valinta |
| 5 | Mallin arviointi |
| | A Mittareiden arvojen määrittäminen vertailumallille ja kehitetylle mallille |
| | B Ajallisen, maantieteellisen ja/tai domain siirrettävyyden tutkiminen |
| | C Mallin potentiaalin määrittäminen aineiston koon suhteen |
| | D Mallin käytön hyötyjen ja riskien arviointi |
| | E Kokonaisarvio perustuen analyysin tuloksiin ja asiantuntijoiden arvioon sekä päätös jatkosta |
| 6 | Mallin integrointi |